Rapid, Precise and Reproducible Binding Affinity Prediction: Applications in Drug Discovery

Srdan Jovanovic

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Chemistry

University College London

I, Srdan Jovanovic, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Signature: ..... Date: .....

The whole of science is nothing more
than a refinement of everyday thinking.

ALBERT EINSTEIN

Abstract

As we move towards an era of personalised medicine, the identification of lead compounds requires years of research and considerable financial backing, in the development of targeted therapies for cancer. We use molecular modelling and simulation to screen a library of active compounds, and understand the ligand-protein interaction at the molecular level in appropriate protein targets, in a bid to identify the most active lead drug candidates.

In recent times, good progress has been made in accurately predicting binding affinities for drug candidates. Advances in high-performance computation (HPC), mean it is now possible to run a larger number of calculations in parallel, paving the way for multiple replica simulations from which binding affinities are obtained. This, then, allows for a tighter control of errors and in turn, a higher confidence in the binding affinity predictions.

Here, we present ESMACS (Enhanced Sampling of Molecular dynamics with Approximation of Continuum Solvent) and TIES (Thermodynamic Integration with Enhanced Sampling); a new framework from which binding affinities are calculated. ESMACS performs 25 replica simulations of the same ligand-receptor system with the only difference being the initial momentum of each atom. From this ensemble of trajectories, an extended MMPBSA (Molecular Mechanics Poisson-Boltzmann Surface Area) free energy method is employed. The TIES protocol constitutes 5 replicas simulations per lambda state followed by the integration of the potential

derivatives of each lambda state, generating a relative binding affinity. This is all tied together using the BAC (Binding Affinity Calculator) which automates the ESMACS and TIES workflow.

ESMACS and TIES, given suitable access to HPC resources, can compute binding affinities in a matter of hours on a supercomputer; the size of such machines therefore means that we can reach the industrial scale of demand necessary to impact drug discovery programmes.

Acknowledgements

I would firstly like to acknowledge, and thank, my supervisor, Professor Peter V. Coveney, for his valuable comments, remarks and engagement that has challenged me intellectually, and allowed me to enjoy my PhD experience. Additionally, I am grateful to Dr. Shunzhou Wan, Dr. David Wright, and PhD colleague and friend, Agastya Bhati, who has been an immediate, positive and vital source of support whenever I needed it.

Thanks also goes to the Centre for Computational Science group, who have made joining the team easy, and are vital in creating a positive working environment, which allows me to perform my research with confidence.

I would like to acknowledge The Engineering and Physical Sciences Research Council (EPSRC), for supporting me financially throughout my studies. An acknowledgement also goes to ARCHER, UK's national High Performance Computing Service, funded by the Office of Science and Technology through EPSRC's High-End Computing Programme; HPC facilities of STFC Hartree Centre; and Liebniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, for all the work presented in this thesis. A final acknowledgement goes to the Qatar National Research Fund (QNRF) and collaborators at Hamad Medical Corporation (HMC) and Carnegie Mellon University – Qatar (CMU-Q).

Thanks to University College London, Department of Chemistry, for offering me the facilities to realise my full potential.

The completion of this thesis would not be possible without the relentless encouragement from my partner, Djana, who has motivated and inspired me throughout the difficult moments.

My deepest and most sincere acknowledgement goes to my parents, Ned and Sabina, and brother, Dado, for their unwavering love, support and commitment throughout my life.

Contents

List of Figures

List of Tables

In memory of my grandfather, Seid Hadžiomerović.

Chapter 1

A brief introduction to biomolecular simulation and applications in drug discovery

A great deal can be learned about biomolecules, such as proteins, through experiment. Let us consider a vial, which is filled with a solution containing a protein and a chemical compound which interacts and binds to this protein (a drug candidate). The solution contains approximately 10^{24} molecules which are interacting with one another. Any measurement that is taken with respect to this sample, returns a macroscopic understanding of the interactions between protein and chemical compound. For instance, if we determine the binding free energy of this sample, this is the average binding free energy of all 10^{24} molecules in the sample. In this context, an experiment is performed using a macroscopic sample that contains an extremely large number of atoms or molecules that sample an enormous number of conformations. It is from this macroscopic ensemble that we are able to gain a physical understanding of drug binding.

However, such experiments tell us almost nothing at the microscopic level. For example, we may be able to determine strength of binding of our hypothetical

drug candidate to the protein through a series of experimental measurements, but we do not know the mechanism of binding on the atomistic level. In other words, how do the atoms of the respective entities interact which subsequently define its binding strength? Detecting this atomic interaction on short timescales in which they occur is virtually impossible using experimental approaches.

X-ray crystallography is an experimental technique within the structural biology domain, that can give us a detailed description of proteins with an atomic resolution. As a result, we know with good confidence, the structure of different types of proteins, from which we can predict the behaviour, such as protein folding and conformational changes. However, large biomolecules are highly dynamic and their motions are critical in all biological processes. In short, it is not entirely possible to infer the dynamics of a protein from a static crystal structure. This can be compared to a photograph of a football match; we can gain exceedingly limited information from this snapshot: to gain a true understanding of the football match, we need to see a video recording [1].

1.1 Simulating biomolecules

This refreshingly simple statement from Richard Feynman [2], neatly summarises the concept of biomolecular simulation: "Everything that living things do can be understood in terms of the jigglings and wigglings of atoms". The central goal is to take static crystal structures, and 'bring them to life' through the use of sophisticated computational techniques to allow us to understanding biological processes on the atomic level. Simulations based on fundamental principles of physics gives us the opportunity to bridge the gap between the structural understanding of proteins, obtained from crystal structures, and the dynamic behaviour of proteins responsible for biological activity.

The increasing power, and parallelism, of computer hardware, coupled with more sophisticated approaches toward simulation and consequent techniques to analyse

simulation trajectories, has meant that biomolecular simulations now play an important part in our understanding of biology [1]. Macroscopic experiments are often accompanied with some sort of atomistic simulation. In fact, rational drug design is a discipline where simulations are used to test hypothesis and interpret experimental data [3]. A good example of this is quantitative structure-activity relationship (QSAR) approaches, where experimental data associated with drug binding is rationalised and interpreted by atomistic biomolecular simulation, and often used to drive forward drug design projects. Artificial Intelligence (AI) and machine learning (ML) approaches are also used to assist chemists in screening large chemical databases. AI and ML technology is self-learning and thus has an ability to learn and improve upon compound database searches on its own. QSAR methods require explicit parameters that are inputted by chemists.

Interactions between proteins and small molecules, such as our hypothetical drug candidate at the beginning of this chapter, are integral in a vast number of biological processes. We turn our attention to the importance of drug binding in the treatment of disease, and hence the use of small molecules as therapeutic drugs. The strength of binding, or binding affinity, of a drug to its protein target, is a physical quantity that is most often computed when designing drugs. Computing this quantity in a rapid and reliable manner has piqued the interest of pharmaceutical companies. Biomolecular simulation and subsequent binding affinity computation, has shown promise in the drug discovery domain as a tool to lower costs and expedite drug design programmes.

1.2 Drug discovery in the pharmaceutical industry

Pharmaceutical companies are being put under increasing pressure, from stake holders and governments alike, to discover and develop cost effective drugs within sustainable financial means. Add to this the increased regulatory hurdles that

need to be cleared, and the impending patent cliff that awaits every new drug, the productivity within the drug discovery and development branch, feels a strong headwind [4, 5].

1.2.1 Growing costs of drug discovery pipelines

Fig. 1.1 displays a simplified representation, of the drug discovery and development pipeline: from a library of millions of compounds, to a final drug being launched into the market [4]. However, achieving this requires a vast amount of time and huge financial backing. DiMasi and colleagues have estimated research and development (R & D) costs since the turn of the century [6]. This was compared to prior studies within a similar time frame [7, 8, 9]. They found that in the last 15 years, the cost per new drug is estimated to be approximately $2.6 billion (Fig. 1.2, [6]). This is nearly a 3-fold increase in cost compared to the '1990s - mid 2000s' time window. In fact it has been growing steadily since the 1970s; the cost increased by a factor of 2.31 from the first to the second time window, and 2.53 times from the second to the third time window. This is a clear indication of the rising costs of drug

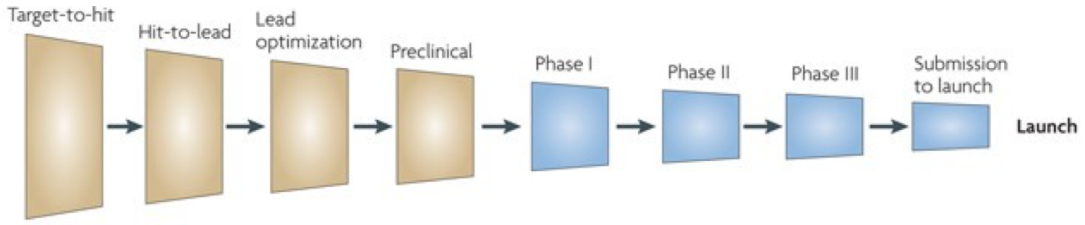


Figure 1.1: *The simplified schematic represents the various stages through the drug discovery and development pipeline. The stages coloured in orange are loosely termed as 'drug discovery' stages, where the blue stages are termed 'development' phases. Drug discovery stages are often described as programmes that involve the synthesis and identification of pharmaceutically active compounds which are subsequently tested in animal models. This leads to the clinical development phases where human testing is introduced in three stages. This schematic does not show the true complexity of this process, as many of the stages in this pipeline often overlap. In additional, pharmaceutical companies will submit investigational drugs that run in parallel with lead active compounds.*

discovery and development costs in pharma.

Both hit-to-lead and lead optimisation stages could benefit from more efficient approaches and a reduction in costs. Hit-to-lead research, in particular, accounts for 6% of the total cost throughout the pipeline, which equates to $166 million, and usually takes 1.5 years, per new molecular entity (NME [4]). In the pharmaceutical industry, $414 million, and 2 years are spent, per NME, to take a drug to launch – financially, that accounts for 17% of the entire process.

Another interesting conclusion is that the success rate of drugs has decreased by 10% since the last study [6], there are several possible reasons for this. Firstly, regulatory authorities, in recent times, have become more stringent, so lead candidates that may have been approved previously are now rejected. Secondly, there has been a shift in focus towards drug discovery in areas of unmet medical need [5], where scientific knowledge is comparatively underdeveloped, or the science is

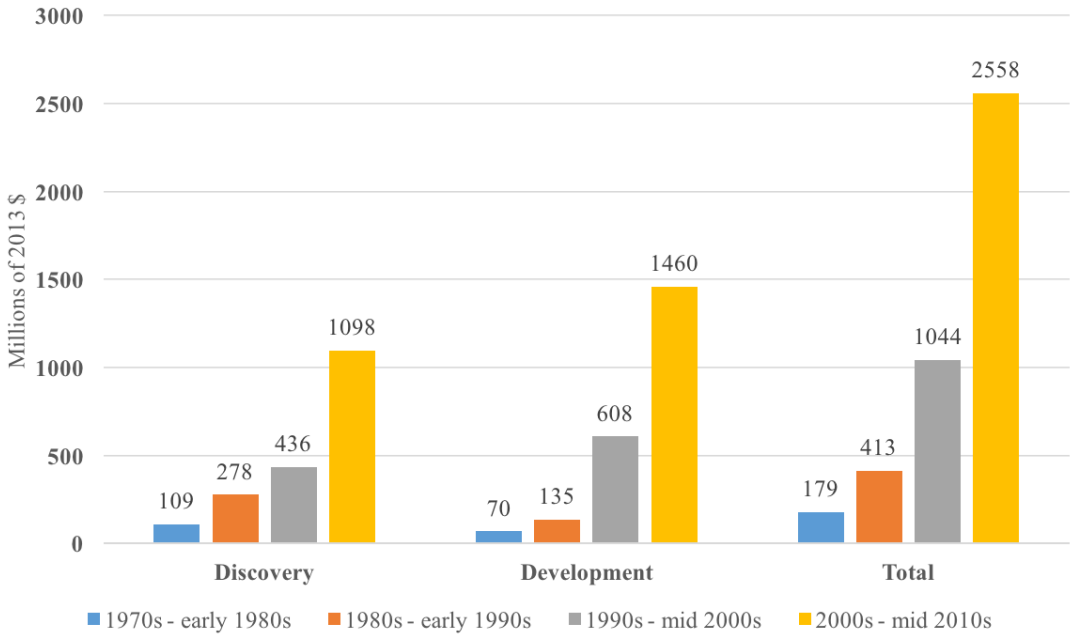


Figure 1.2: *Trends in costs for the discovery and development phases in the pharmaceutical pipeline, and the combined amount, over a period of circa 45 years. A steady increase in costs for both sectors of the pipelines is witnessed.*

difficult, leading to high percentage of failing drug candidates. Lastly, the influence of molecular biology and subsequently genomic sciences, has led to the detection of increasing number of drug targets. Although this is a vital development in the field of drug discovery, there are cases where some drug targets have been poorly validated leading to the more drug candidates being developed, with little knowledge of its success rates [10, 11].

1.2.2 Reliability of research

Needless to say, in an industry where the health of patients is in question, methodologies need to be reliable and reproducible. This would seem obvious, but recent publications [12, 13] have shown that scientific research has frequently produced unreliable, and irreproducible results, leading to false claims. The consequence of this is that subsequent research efforts that aim to build upon these irreproducible results, are also worthless. This has direct financial implications to pharmaceutical companies – not withstanding the pursuit of scientific integrity – because potentially millions of dollars could be invested, based on scientific results that, at the time, seemed plausible, but turn out to be false. In fact, one of the leading biopharmaceutical companies reviewed 53 "landmark" publications and of these, only 6 (11%) of the results were successfully reproduced [12]. A similar case was reported which involved another major pharmaceutical company, where only 25% of publications could be validated [14]. This level of unreliability in methodologies is unacceptable, especially in an industry where reliable and reproducible results are of utmost importance.

1.2.3 Applications in early stage drug discovery

Hit-to-lead is a period during early stage drug discovery where small molecules, which exhibit biochemical activity in a particular biological target (known as a hit), are evaluated, to identify a lead compound which then undergoes further drug

development [15]. Quantifying this biochemical association between a molecule and biological target, known as the binding affinity, is essential in the pharmaceutical industry, in the pursuit of identifying lead drug candidates. The hit-to-lead stage usually means reducing hundreds of active compounds down to a handful of the most promising drug candidates. The distinction between the best performing drug candidates and the remainder is usually achieved by experimentally calculating the binding affinity.

Lead optimization proceeds hit-to-lead, where the lead candidates from the latter stage, are tested to ascertain their pharmacokinetic (PK) – absorption, distribution, metabolism and excretion) – and pharmacodynamic (PD) or drug-receptor interactions profile. At this stage, the chemical structure of the lead candidate is confirmed, however to improve upon already comparatively good PK/PD profiles, small deviations are made to the structures, to induce, for instance, selectivity to a particular receptor, or a better metabolism profile.

However, on an industrial scale, experimental techniques tend to be either labourious, imprecise or costly (or all). In additional to this, medicinal chemists spend a vast amount of time and resources on synthesising, purifying and validating drug candidates.

1.2.4 Personalised medicine

The personalised medicine domain will also benefit from improved drug design methods. A clinician will have an array of medicines, which can then be used as bespoke treatments for individual patients, based on their genetic profile. Predicitive tools to assist clinicians would make clinical decision-making more efficient by administering the right therapy for each individual patient. The use of molecular dynamics (MD) smulations from which drug binding affinities are estimated has already shown promise in several different biological systems [16, 17, 18, 19].

1.2.5 High-performance and cloud computing resources

Computational drug discovery, along with concurrent developments in structural biology and genomics, has grown rapidly not least in part to the advancements in high-performance computers (HPCs) and cloud computing. HPCs offer the capabilities of parallelism, which allow computational scientists to run a number of simulations concurrently across a large number of processors. Alternatively, cloud computing allows users to run computational drug discovery applications 'on-demand' using remote resources. Section 3.4 describes in more detail the costs and related issues associated with HPC facilities and the use of cloud computing.

1.2.6 Concluding remarks

Over approximately the last four decades, biomolecular simulation has changed from picosecond simulations of crude macromolecular models, *in vacuo*, to millisecond simulations of complex and heterogeneous biological systems consisting of millions of atoms. The development over this time has been hugely promising. However, it is only now that such approaches are being integrated into the pharmaceutical industry.

The pharmaceutical industry faces huge financial challenges due to rising costs and increased regulatory hurdles. Additionally, there is room for improvement in terms of productivity: in the last 15 years, there has been an increase in failed new drug applications compared to the previous 15 years [6]. A closer inspection of the hit-to-lead and lead optimiziation stage of the drug discovery pipeline has shown that much of the efforts are exerted here, both financially and temporally. There are many failures in clinical trials too: these 'late' failures are particularly expensive and need to be avoided. In tandem with more efficient drug discovery strategies, reproducible results are essential, bringing reliability and confidence to the research results obtained.

Hence, there is unmet need in the pharmaceutical industry for financially viable methods that can achieve reliable results, time and again. Computational developments that allow for predictive modelling of biomolecular systems, together with emergence of multi peta-scale computational resources, have opened up an opportunity to drive such costs down. Computer-aided rational drug design is a division within drug discovery which could directly benefit from such advances. Molecular modelling techniques and the correct free energy calculation method can be incorporated to expedite this process of identifying a lead, or leads. The introduction of molecular modelling tools within this space has the potential to assist in the current experimental techniques, and as such, reduce the time and cost.

Computational Biomedicine (CompBioMed) is a Centre of Excellence in Computational Biomedicine that aims to advance the role of computional based modelling and simulation within biomedicine [20]. This will be achieved through the use of international HPC resources, and development of software tools which are capable of delivering high fidelity modelling and simulation of the human body. The innovative modelling and simulation techniques will be of key interest to industrial researchers, HPC manufactures and scientific software developers.

Chapter 2

Theory

This chapter will outline the fundamental theoretical concepts that embody the topic of binding affinity predictions. This begins with a kinetic and thermodynamic definition of the processes involved in the binding of a drug to its target protein. The ability to predict binding affinities using computational means requires understanding of molecular simulation, which links the macroscopic scale of experiment to the microscopic scale of simulation. A description of this connection, through the use of statistical mechanics is also presented here. Finally, theoretical and computational techniques used to quantify the binding affinity are explained.

2.1 What is a binding affinity?

In pharmacology, a ligand is a therapeutic drug that inhibits the biological responce of its target which is usually a protein. This type of ligand is often termed an antagonist. This is achieved through the process of binding, where the ligand will attach onto the protein at either the active (orthosteric) site, or elsewhere on the protein (allosteric site) [21]. The strength of this binding, through the various interactions made between the ligand and protein determine the binding affinity of the drug to the protein. The affinity alone does not give us a measure of the performance of the drug [21]. A drug may have a very high binding affinity, but low

efficacy. Thus, efficacy is the measure used to describe the action of the drug once it is bound to the protein. Binding affinity and efficacy determine the effectiveness of a therapeutic drug to inhibit the biological response of a protein – this term is called the potency [21]. The following sections give a theoretical explanation of the binding affinity, namely the kinetic and thermodynamic considerations involved in the protein-ligand binding process. A more detailed description of these topics can be found in the following sources [21, 22, 23].

2.1.1 Protein-ligand binding kinetics

A ligand binds to a protein to form a complex. This process can be reversible, where the interactions between ligand and protein are non-covalent, or it can be irreversible, where the interactions are covalent. In reversible binding, a protein and ligand are also able to dissociate and a ligand is subsequently free to associate with a different protein molecule. Drugs, in most cases, are designed to bind reversibly, and so are attention will be directed to this mechanism of binding.

Protein-ligand kinetics describes the association between the two entities, and the rate at which they bind to each other. For example, when a protein and ligand are mixed in solution, with fixed concentrations, the rate of association between the two can be described as such:

$$P + L \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} PL \tag{2.1}$$

where P is the protein, L is the ligand, and PL is the protein-ligand complex. The rate constants, k_1 $(M^{-1} \cdot s^{-1})$ and k_{-1} (s^{-1}), are the kinetic rate constants that represent the forward reaction of ligand binding or association, and the reverse reaction of the ligand unbinding from the protein, termed dissociation. At equilibrium, the rate of the forward reaction of ligand binding $(P + L \rightarrow PL)$, and the reverse reaction of unbinding $(PL \rightarrow P + L)$, are equal. This can be represented

like so:

$$k_1[P][L] = k_{-1}[PL] \tag{2.2}$$

where the terms in square brackets, $[\cdots]$, corresponds to the concentration of each entity, at equilibrium.

At equilibrium, then, the association constant, K_a (M^{-1}), is equal to the reciprocal of the dissociation constant, k_d (M). In other words, a fast k_1 and a slow k_{-1} yield a high association, and low dissociation constant, which results in a high binding affinity:

$$K_a = \frac{k_1}{k_{-1}} = \frac{[PL]}{[P][L]} = \frac{1}{K_d} \tag{2.3a}$$

$$\frac{[PL]}{[P]} = K_a[L] \tag{2.3b}$$

Rearranging Eqn. 2.3a to Eqn. 2.3b, shows that increasing the ligand concentration, gives rise to a large number of protein-ligand complexes. Further, a higher concentration of complexes, means a larger K_a value.

2.1.2 A thermodynamic representation

Alternatively, one can understand protein-ligand binding by employing a thermodynamic representation. The driving forces that characterise the binding of ligand to a protein, in solution, is explained through the energy exchange between the three components (P, L and PL). These driving forces can be quantified as the Gibbs free energy (G).

The Gibbs free energy, G, is a thermodynamic potential that measures the maximum reversible work that can be performed by a thermodynamic system at con-

stant pressure and temperature (isobaric-isothermal or NPT system). The change in Gibbs free energy (ΔG) can give us information about the spontaneous process of protein-ligand binding. That is, the thermodynamic system, protein and ligand in solution, needs to have a negative ΔG for protein-ligand association to occur. In other words, the difference in G between the reactants and products requires a negative value. This process will occur until the reaction reaches equilibrium, and thus ΔG becomes zero:

$$\Delta G = G(PL) - G(P) - G(L) \tag{2.4}$$

The relationship between K_a and the standard binding free energy (ΔG°) can be described according to Eqn. 2.5:

$$\Delta G^\circ = -RTlnK_a \tag{2.5}$$

where the free energy is obtained under standard conditions: 1 atm of pressure, temperature of 298 K and 1 M concentrations of protein and ligand. The gas constant, R, is 1.987 $cal \cdot K^{-1}mol^{-1}$, and T is the temperature. We see here that as the K_a value increases, ΔG° becomes more negative, resulting in a higher binding affinity. Hence, the kinetic rate constants, k_1 and k_{-1} are connected to the thermodynamic property, ΔG°.

$$\Delta G = \Delta G^\circ + RTlnQ \tag{2.6}$$

It was mentioned earlier that when a spontaneous thermodynamic reaction in isobaric-isothermal conditions reaches equilibrium, ΔG becomes zero. This is seen in Eqn. 2.6 which describes the binding free energy, ΔG, during protein-ligand binding, not limited to standard conditions. The reaction quotient, Q, is the ra-

tio of the concentration of the protein-ligand complex ($[PL]$), and the product of the concentration of the free protein and ligand ($[P][L]$). When a reaction is in equilibrium, $K_a = Q$ and ΔG is zero.

Eqn. 2.7a describes ΔG as enthalpic and entropic contributions to binding:

$$\Delta G = \Delta H - T\Delta S \tag{2.7a}$$

$$\Delta G = (\Delta U + p\Delta V) - T\Delta S \tag{2.7b}$$

where ΔH is the change in enthalpy, and $T\Delta S$ is the change in entropy as a function of temperature, in Kelvin. Enthalpy is the total energy of the thermodynamic system. This includes the internal energies of the protein and ligand (U), and the energy needed to displace the surroundings, which is a product of the volume (V) and pressure (p) of the system (Eqn. 2.7b). Protein-ligand binding is an exothermic process, where the non-covalent bonds that are formed are energetically favourable. Thus, the change in enthalpy (ΔH) is the change in the energies between the reactants and products, where a spontaneous binding process yields a negative ΔH value.

Entropy is the overall disorder of the atoms within a system. The process of ligand-binding sees two molecules, the free ligand and protein, form one protein-ligand complex, and by definition, the entropy is higher when the components are free, than when the components are complexed. Therefore $T\Delta S$ is a positive value in spontaneous protein-ligand binding, and acts as an entropic cost to the ΔG value.

$$\Delta S = \Delta S_{solv} + \Delta S_{conf} + \Delta S_{r/t} \tag{2.8}$$

The total entropy change can be segmented into three terms, seen in Eqn. 2.8. ΔS_{solv} is the entropy change due to solvation. Although the formation of a cavity

in a solution is entropically unfavourable, as the system becomes more ordered, the entry of the protein-ligand complex into this cavity is entropically favourable. The net value of $T\Delta S$ is positive. The change in conformational entropy ΔS_{conf} between free ligand and protein, and protein-ligand complex can be either positive or negative, and depends on the protein and ligand in question. Finally, $\Delta S_{r/t}$ is the change in entropy of the rotational and translation degrees of freedom as a result of ligand binding. This contributes unfavourably to entropy there is a loss in rotational and translation freedom in the protein and ligand, in the complex formed.

2.1.3 Kinetics of enzyme action

The function of an enzyme is to increase the rate of a reaction, without any change to the enzyme concentration itself. The rate of reaction is determined by the activation energy: this is the highest energy state that needs to be overcome by the reactants to form the product. Enzymes lower the activation energy, and so the rate of reaction increases. This can be illustrated using the Michaelis-Menten equation [24] which relates the reaction rate to substrate concentration.

The activation energy is equivalent to the binding free energy between the enzyme and the enzyme-substrate complex. Enzyme kinetics can be described like so:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} E + P \tag{2.9}$$

where k_1 and k_{-1} are the association and dissociation constants for the enzyme and substrate, and k_2 is the rate constant for the catalysed reaction. Here we assume, for simplicity, that the catalysed reaction is irreversible. This is because it is highly unlikely, due to the thermodynamic stability of the enzyme and product, that it will follow the reverse reaction to give the enzyme-substrate complex. The rate of reaction (V) is the rate at which product is formed is:

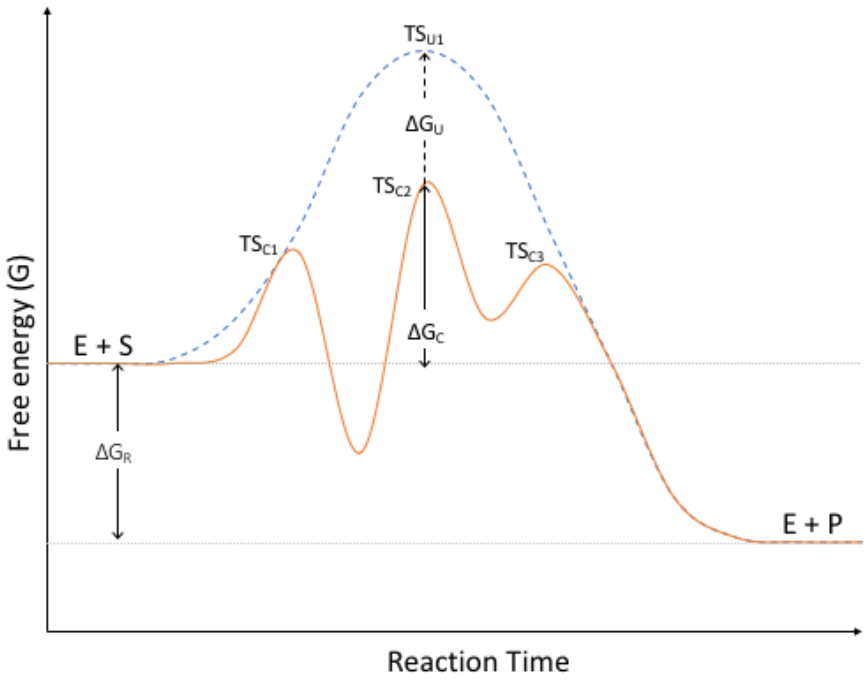


Figure 2.1: *A schematic of the course of a reaction for processes that are catalysed (orange) and uncatalysed (blue). A catalysed reaction is represented by a smaller free energy barrier (ΔG_C) and so the reaction can proceed at a faster rate. An uncatalysed reaction has a higher free energy barrier (ΔG_U). In the catalysed reaction, several transition states are formed, which eventually results in the formation of product. The change in free energy of the reaction (ΔG_R) is the difference in free energy between the reactants and products.*

$$V = \frac{d[P]}{dt} = k_2[ES] \tag{2.10}$$

where $[\cdots]$ indicates the concentration of the product or enzyme-substrate complex. The rate at which the concentration of the enzyme-substrate complex changes is equivalent to the rate of its association, minus the rate of its dissociation as shown in Eqn. 2.11:

$$\frac{d[ES]}{dt} = k_1[E][S] - (k_{-1} + k_2)[ES] \tag{2.11}$$

Here, we introduce the idea that the total number of enzyme available ($[E]_T$), is

equal to the free enzyme plus the enzyme-substrate complex, which can be formulated as $[E]_T = [E] + [ES]$. Taking this into account we get the Eqn. 2.12a, which is further simplified by making two assumptions. The steady-state assumption says that the rate of enzyme-substrate complex formation is equal to the rate of formation of the product and the dissociation back to free substrate and enzyme (Eqn. 2.12b). The second assumption, shown in Eqn 2.12c, also holds true, as the concentration of the enzyme-substrate complex is much larger than the term on the left side, for the majority of the reaction time.

$$\frac{\frac{d[ES]}{dt}}{k_1[S] + k_{-1} + k_2} + [ES] = \frac{k_1[E]_T[S]}{k_1[S] + k_{-1} + k_2} \tag{2.12a}$$

$$\frac{d[ES]}{dt} = 0 \tag{2.12b}$$

$$\frac{\frac{d[ES]}{dt}}{k_1[S] + k_{-1} + k_2} \ll [ES] \tag{2.12c}$$

Thus, the Michaelis-Menten equation is derived from Eqn. 2.11 and 2.12a. The Michaelis constant K_m is substituted, for $\frac{k_{-1}+k_2}{k_1}$ under the equilibrium approximation, $k_d \ll k_{cat}$:

$$V = k_2[ES] = \frac{k_2[E]_T[S]}{[S] + K_m} \tag{2.13}$$

where V_{max} is the maximum rate of reaction when all enzyme molecules are fully occupied by the substrate. When the rate of reaction is at the maximum speed, in other words when $V_0 = V_{max}$, then the total enzymatic concentration, $[E]_T$, is the same as the concentration of the enzyme-substrate, since all the enzymes are occupied with substrate ($[E]_T = [ES]$). Substituting this into Eqn. 2.13 gives the simplified Michael-Menten equation:

$$V = \frac{V_{max}[S]}{[S] + K_m} \tag{2.14}$$

But what is the importance of K_m? Consider that K_m equals the substrate concentration, $[S]$. Then, substituting this into the Michaelis-Menten equation, we see that K_m is the substrate concentration where the rate of product formation is half of the maximum reaction rate, V_{max}.

$$V = \frac{1}{2}V_{max} \tag{2.15}$$

So the lower the K_m, the better an enzyme can function when substrate concentrations are limited.

2.1.4 Enzyme inhibition

Reversible inhibitors fall in to three categories: competitive, uncompetitive and non-competitive inhibitors. Competitive inhibitors bind to the free enzyme and slow down the rate of enzyme-substrate formation. Thus, such inhibitors compete with the substrate at the binding side. Uncompetitive inhibitors, conversely, bind to the enzyme-substrate complex and do not directly compete for the binding site of the substrate. Intuitively, uncompetitive inhibitors binding elsewhere on the enzyme to inhibit the formation of product. Non-competitive inhibitors are also known as "mixed" inhibitors, because they can bind to the free enzyme or the enzyme-substrate complex, to inhibit the rate of product formation.

Properties associated with enzyme inhibition can be described effectively using Lineweaver-Burk plots [25], which can be derived by taking the inverse of both sides of the Michaelis-Menten equation (Eqn. 2.14):

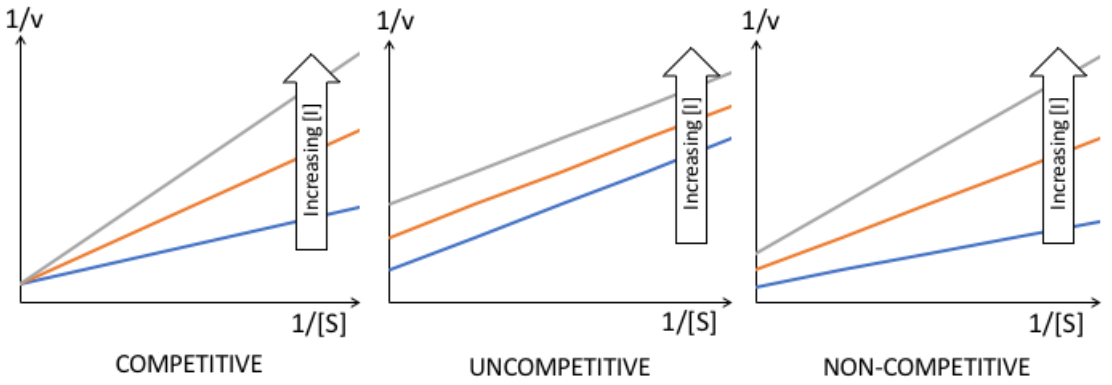


Figure 2.2: *Lineweaver-Burke plots illustrating the effects of increasing inhibitor concentration ([I]), on maximum reaction rate (V_{max}) and the binding affinity of the substrate to the enzyme K_m. The blue and grey lines are low and high [I], respectively. The plots are conceptual and are not an accurate representation.*

$$\frac{1}{V} = \frac{K_m}{V_{max}} \cdot \frac{1}{[S]} + \frac{1}{V_{max}} \tag{2.16}$$

where $1/V$ is the dependent variable, K_m/V_{max} represents the slope of the line function, $1/[S]$ is the dependent variable, and $1/V_{max}$ represents the y-axis intercept. Using this, we can determine the activity of each type of inhibitor, with respect to K_m and V_{max}

As the concentration of competitive inhibitor is increased, the slope of the lines increases too, but the y-intercept is the same. This means that an increased concentration of competitive inhibitor will increase the K_m, resulting higher inhibition at low substrate concentrations. However, the V_{max} stays unchanged (same y-intercept), and so competitive inhibitors do not effect the maximum reaction rate of the enzyme. In other words, at high concentrations of substrate, the inhibitor loses effect.

An increase in uncompetitive inhibitors sees a decrease in V_{max}, because the y-intercept ($1/V_{max}$) is increasing, and no change in K_m. So at low concentrations of substrate, an increase in uncompetitive inhibitor does not effect the enzyme activ-

ity, but at high substrate concentrations, uncompetitive inhibitors remain effective.

Finally, increasing the concentration of non-competitive inhibitors, we see no change in K_m, and a decrease in V_{max}. So these inhibitors are still effective at high concentrations of substrate, but the enzyme activity is unchanged at low concentrations.

As in Eqn. 2.9, we are able to depict competitive binding in a similar fashion:

$$E + S + I \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} E + P + I \tag{2.17a}$$

$$E + S + I \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} EI + S \tag{2.17b}$$

where the notation is consistent with equation 2.9, but includes the inhibitor, I, and the rate constants for for the association k_3 and dissociation k_{-3} of the enzyme-inhibitor complex. Following the same derivation of the Michaelis-Menten equation (Eqn. 2.14) we obtain:

$$V = \frac{V_{max}[S]}{K_m(1 + \frac{[I]}{K_i}) + [S]} \tag{2.18a}$$

$$K_i = \frac{k_{-2}}{k_2} \tag{2.18b}$$

where inhibition constant, K_i, is the equal to the disassociation constant (Eqn. 2.18b). The introduction of Ki is important, because now we can quantify the binding affinity of an inhibitor to an enzyme. A low K_i value, corresponds to a high rate of enzyme-inhibitor formation (k_3) and thus a higher binding affinity.

2.2 Experimental binding affinities

Testing the interaction between two molecules, is one of the most common experiments in biochemistry and molecular biology. For this reason, there are a vast amount of different techniques used to measure these interactions [26]. Our interest is focused on the interaction of a ligand with a protein, and in this domain, there are several techniques that are preferred. Binding affinity experiments fall into two categories. Direct methods, which measure the actual concentrations within the sample, and indirect methods, which measures a signal from an external source, and is subsequently converted into a binding affinity value. The most common methods are described below.

2.2.1 Isothermal titration calorimetry

The gold standard of experimental binding affinity measurements is isothermal titatration calorimetry (ITC, [27]). This quantitative thermodynamic approach is the only protocol that can directly measure physical properties, like ΔG, ΔH and ΔS, in addition to the K_a constant. In addition, when ITC experiments are conducted on the same protein-ligand system, with variations in temperature, the heat capacity (ΔC_p) can also be measured.

First, the ligand of interest is injected, in small aliquots, into a solution containing the protein. The aliquots are precisely titrated into the sample cell which triggers a change in temperature relative to a reference cell. Then, the cell heater responds by heating or cooling, depending on whether the reaction is exothermic or endothermic, to return the sample cell to base temperature. The power applied by the cell heater to the sample cell after each titration of ligand, can be converted into the heat produced by the cell.

$$q_i = v\Delta H \Delta[L]_i \tag{2.19}$$

The energy required decreases as the titration proceeds, as there is less protein on to which the ligand can bind. Finally, the change in heat over the entire reaction time can be used to calculate ΔH directly, using Eqn. 2.19, where v is the volume of the cell, q_i is the heat generated for each aliqout of ligand titrated into the sample, and $[L]_i$ is the concentration of ligand for each aliquot. The K_a can also be calculated from the total amount of heat produced which can, in turn, be used to calculate ΔG and and ΔS via Eqns. 2.5 and 2.7a.

2.2.2 Fluorescence spectroscopy techniques

This group of techniques measure the rate of product formation, or the rate at which the substrate is associating with the enzyme, through the use of fluorescence.

The general idea of fluorescence-based experiments is to label either the ligand or substrate with a marker that has fluorescent properties. A fluorometer detects the strength of this signal. The reaction between the competing ligand and substrate is initiated, resulting in a change of strength in fluorescence signal. For example, a protein and substrate are allowed to interact where initially, there is no fluorescent signal. Upon the addition of a known concentration of a competitive inhibitor, which has been labelled with a fluorescent marker, the strength of the fluorescent signal indicates the concentration of inhibitor occupying the binding site.

The two main techniques are fluorescence polarisation (FP, [28, 29]) and fluorescence resonance energy transfer (FRET, [30, 31]). FP uses the idea of polarised fluorescence emission which becomes unpolarised faster in the bound state, than the unbound state. Here the inhibitor is excited using polarisable light, and the speed at which it becomes unpolarised indicates the amount of ligand bound to the enzyme. FRET requires a double labelling of the enzyme and inhibitor. When apart, the labels do not emit fluorescent light, but when the inhibitor binds to the enzyme, the energy transfer between the two labelled molecules emits a fluorescent signal. The strength of this signal corresponds to the amount of inhibit occupying

the enzyme.

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_m}} \tag{2.20}$$

The results of these techniques are usually reported as the concentration of inhibitor required to reduce the enzyme activity by half (IC_{50}). This value can be related to the K_i due to the reasoning that at low values of $[S]$, the K_i equals the IC_{50}. The Cheng-Prussoff equation [32], in Eqn 2.20, demonstrates this relationship.

2.2.3 Surface plasmon resonance

Surface plasmon resonance (SPR, [33]) is an optical based technique that measures the change in refractive index due to ligand association and dissociation. Protein molecules are immobilsed on a sensor surface and the analyte molecules, that is the inhibitor, is injected onto the surface allowing for inhibitor association. This association of inhibitor to the immobilised protein is accompanied by an increase in refractive index, yielding the rate of ligand association, k_1. After some time, a solution that dissociates the ligand (usually the sample buffer) from the protein is introduced, and the refractive index is measured, giving rise to the rate of ligand dissociation (k_{-1}). Using Eqn. 2.3, the association constant, K_a, can be deduced.

2.2.4 Experimental error

Experimental techniques are associated with errors due to a number of factors, and the difficulty in controlling them. Here, potential sources of error [34, 35, 36, 37, 28, 38] will be outlined for the approaches that have been explained in section 2.2 .

Although ITC is highly sensitive, the detection of heat from protein-ligand association, for systems that report very low enthalpies, is difficult to measure accu-

rately. Similarly, ITC is insensitive in cases where the kinetic rate of reaction is very slow. In addition, the technique requires a large sample size, and so proteins and inhibitors that are difficult to prepare, or expensive in large quantities, are not suited for this method. The experiment itself is time-consuming and low-throughput and for this reason is would not be well placed in a drug discovery programme where a large amount of inhibitors need to be tested in a relatively short space of time. Further, the laborious and intricate nature of ITC means that replicates are rarely performed and so there is often a lack of error bars associated with the measurement.

SPR has the benefit of real-time analysis of the refractive index and so a clear understanding of the reaction kinetics is available. However, a large draw-back of this method is the immobilisation of the protein, which constrains the protein conformational, translation and rotation degrees of freedom. As a result, the k_a constant is not representative of the actual ligand binding process.

Fluoresence-based assays are commonly used as they are relatively inexpensive, high-throughput, and require small sample sizes. For this reason, they are suited for drug discovery programmes as the protocol can be automated. On the other hand, factors such as light-scattering and auto-fluorescence can alter the true signal that is detected. Further, labelling the ligands or enzymes alters the binding behaviour, similarly to what is experienced in SPR.

2.3 Computational binding affinities

An alternative approach to measuring binding affinities is to predict them using computational means. This is done by using a theoretical framework to define the binding free energy, and the calculations associated with this are then executed using computers. The benefits of this approach are numerous. Firstly, there is no need for reagents, solvents, chemicals and large laboratory space to generate binding affinities. It can all be completed in the comfort of one's office desk.

Additionally, the process of synthesising new and complex small molecules, is a laborious and expensive process, and computational techniques can overcome this problem. These methods also have the benefit of giving us information about the atomistic properties of binding.

The basis of computational prediction of binding affinities is linking the microscopic simulation that is performed on computers, with the macroscopic thermodynamic description of an experiment. Statistical mechanics connects these two scales. The following sections explain how statistical mechanics allows us to run a simulation of a single protein-ligand complex, and achieve comparable binding affinities of an experiment that contains 10^{24} molecules (like the one describes at the beginnig of this thesis). A thorough explanation of the following topics, and more, can be found by accessing the following textbooks. One of these sources provides an introductory explanation of the theory behind a number of computational chemistry approaches [39]. The other focusses more on the principles and applications within molecular modelling [40].

2.3.1 Statistical mechanics

The macroscopic and microscopic system are connected through statistical mechanics, and a key component of understanding this link, is the partition function. At temperature of absolute zero ($0K$), all molecules are in their ground energetic state. However, at any other temperature, there is a probability associated with a molecule being in any particular energy state (relative to the ground state energy). This is described by the Boltzmann factor:

$$P \propto e^{\epsilon/kT} \tag{2.21}$$

where P is the relative probability, ϵ is the energy state of the molecule, T is the temperature, and k is the Boltzmann constant. Eqn. 2.21 demonstrates that,

although there are many more states with high energy than low energy, the probability of finding a molecule with low energy is minimal. A Boltzmann energy distribution plot shows the probability of a molecular being in a particular energy state.

$$q = \sum_{statesi}^{\infty} e^{\epsilon_i/kT} \tag{2.22a}$$

$$P(\epsilon_i) = \frac{e^{\epsilon_i/kT}}{q} \tag{2.22b}$$

The partition function (q), then, is the sum of all possible microstates of a single molecule (Eqn. 2.22a), and can also be viewed as a weighted probability of the molecule being in any particular energy state (Eqn. 2.22b). The partition function q, describes a single molecule that has no interactions (i.e. behaves like an ideal gas), however, a protein and ligand in solution is a system that contains many molecules, which are all interacting with one another. Therefore, a partition function, Q, is calculated by summing over all energy states for the entire system (Eqn. 2.23a) and can also be presented as weighted probability of the system being in a particular energy state (Eqn. 2.23b):

$$Q = \sum_{i}^{\infty} e^{E_i/kT} \tag{2.23a}$$

$$P(E_i) = \frac{e^{E_i/kT}}{Q} \tag{2.23b}$$

The discrete sum of the energies, can also be replaced by an integral of the coordinates ($\mathbf{r}$) and momenta ($\mathbf{p}$) of a system, where $\mathbf{p}$ is the product of velocity and mass, and this known as phase space. This equation becomes important when considering molecular simulation, which will be explained later:

$$Q = \int e^{E(\mathbf{r},\mathbf{p})/kT} d\mathbf{r} d\mathbf{p} \tag{2.24}$$

The importance of the partition function is that it allows us to calculate thermodynamic properties, such as internal energy (U), Helmholtz free energy (H) and Gibbs free energy (G). Note, that U and G is related to Q via a derivative, where A is directly linked to Q:

$$U = kT^2 \Big(\frac{\partial \ln Q}{\partial T} \Big)_V \tag{2.25a}$$

$$A = -kT \ln Q \tag{2.25b}$$

$$G = H - TS = kTV \Big(\frac{\partial \ln Q}{\partial V} \Big)_T - kT \ln Q \tag{2.25c}$$

For small, di- or tri-atomic systems, it is possible to calculate the partition function, q, directly. However, for condensed phases, like a protein-ligand solution, it is not possible to calculate Q by summing over all energy states, or integrating the whole of phase space as the amount of configurations possible is too large. It is possible, though, to estimate Q by generating a representation of the system, where only a portion of phase space is sampled.

2.3.2 Time and ensemble averages

Thermodynamic properties are dependent on the position ($\mathbf{r}$) and momenta ($\mathbf{p}$) of N particles, that make up a system. From a time-dependent MD trajectory, then, a value of a particular thermodynamic property (i.e. G) is defined by the position and moment of N particles at time t. As t tends to zero, G changes as the system changes, and a time average of Gibbs free energy, $\bar{G}$, is obtained like so:

$$\bar{G} = \lim_{t \to \infty} \frac{1}{t} \int_0^t G(\mathbf{r}^N(t), \mathbf{p}^N(t)) dt \tag{2.26}$$

where $\mathbf{r}^N(t)$ and $\mathbf{p}^N(t)$ are the position and momenta of N particles of a system at time t. Eqn. 2.26 shows that as the time of the simulation approaches infinity, the integral approaches the 'true' value of G. To obtain the true value of $\bar{G}$, a molecular simulation must record configurations infinitely, which is clearly not plausible.

$$\langle G \rangle = \iint G(\mathbf{r}^N, \mathbf{p}^N) \rho(\mathbf{r}^N, \mathbf{p}^N) d\mathbf{r}^N d\mathbf{p}^N \tag{2.27}$$

Ensemble averages (Eqn. 2,27) treat each snapshot in time of an evolving system as an individual 'microstate', rather than a 'macrostate', meaning that position and momenta of N particles for each time t are considered independently and simultaneously. Where $\langle G \rangle$ is the ensemble average of Gibbs free energy, and $\rho(\mathbf{r}^N \mathbf{p}^N)$ is the probability density of the ensemble. This was previously shown in Eqn. 2.24, where the energy state (E) can be replaced with the continuous phase space formulation, $\rho(\mathbf{r}^N \mathbf{p}^N)$. The connection between time and ensemble averages is made by applying the ergodic hypothesis, $\langle G \rangle = \bar{G}$.

2.3.3 Statistical uncertainty of binding affinity predictions

We noted that thermodynamic properties, such as U and G are related to the partition function, Q, through a derivative, and conversely, A is directly related to Q. This is also demonstrated by deriving formal expressions for U and A, from the idea that $\partial \ln Q / \partial T = Q^{-1} \partial Q / \partial T$, and then rewriting as an integral using the continuous phase space notation:

$$U = \int E(\mathbf{r}^N, \mathbf{p}^N) P(\mathbf{r}^N, \mathbf{p}^N) d\mathbf{r}^N d\mathbf{p}^N \tag{2.28a}$$

$$A = kT \ln \int e^{E(\mathbf{r}^N, \mathbf{p}^N)/kT} P(\mathbf{r}^N, \mathbf{p}^N) d\mathbf{r}^N d\mathbf{p}^N \tag{2.28b}$$

Here, U is a sum of the weighted probabilities of the system being in a particular energy state, and thus has a linear relationship with $E(\mathbf{r}^N, \mathbf{p}^N)$ (Eqn. 2.28a).

The areas of phase space with a low probability of being sampled, that is high energy configurations, contribute little to the final U value, and so U converges quickly. On the hand, the exponential connection of A with $E(\mathbf{r}^N, \mathbf{p}^N)$ (Eqn. 2.28b), which is synonymous with G, states that the infrequently visited, high energy configurations, contribute significantly to the final thermodynamic value, and so these values do not converge quickly.

Calculating G is challenging, because of the statistical uncertainty surrounding the sampling of a sufficient number of configurations. The statistical uncertainty is proportional to the inverse square root of the number of configurations sampled:

$$\sigma(X) \propto \frac{1}{\sqrt{M}} \tag{2.29}$$

where σ is the statistical uncertainty, and M is the number of configurations sampled. Alternatively stated, if the sample size is increased, then statistical error decreases. However, we have learned that there are limits to simply increasing the sample size, and one must generate a representative sample. Obtaining a representative sample that visits a small area of phase space, frequently, will yield a result that has low statistical error, but high systematic error. In other words, the value will be precise, but inaccurate. Calculating U and other energy properties do not have this problem, as results converge quickly at low-energy regions, but entropic properties, like G are reliant on the whole of phase space, and so obtaining the absolute value of G is impossible. The alternative, though, is to calculate differences in G, which gives rise to a relative change in G.

2.3.4 Ensemble simulations

The most common approach to calculating thermodynamic properties, like the binding free energy of a ligand to a receptor, is by running a 'long' simulation, traversing a part of the energy landscape, and then, using the ergodic hypothe-

sis, claiming that this represents the macroscopic thermodynamic properties of a system [41]. The error with this approach is that the ergodic theorem requires all states to be passed which is beyond our computational capabilities due to the enormity of the number of states available for a biomolecular system.

An alternative approach is to execute more independent simulations of the same ligand-receptor complex. This approach stays true to the theory of statistical mechanics by calculating average values of thermodynamic properties from n replications of ensemble averages. Calculating binding free energies from ensemble simulations recognises that systems in an equilibrium state exhibit dynamics that are chaotic. As a result, binding free energies computed from independent trajectories are described by Gaussian random processes. Thus, it is a stochastic process from which the binding free energy can be obtained statistically conforming to a Gaussian distribution (Fig. 2.3). Ensemble simulations, then, allow us to exploit stochastic calculus by reintepreting many equations, to determine binding free energies. With respect to phase space, ensemble simulations explore a larger area compared with a single trajectory [41]. In addition, the use of replicas allows for reproducible results with error bars, which are scientifically much more valuable. The studies detailed in later chapters embrace ensemble-based approaches, and have been shown to produce rapid, reliable and precise binding affinity predictions for the reasons described above [42, 16, 43, 18, 19].

Some groups have reported accurate approaches to predicting binding free energies [44, 45, 46], but there is no defined automated, ensemble-based workflow that also takes reproducibility and speed into account.

2.4 Generating a trajectory of configurations

Sampling the entirety of phase space, for quite large systems such as protein-ligand complexes in solution, is not possible. Computer simulation techniques, allow us to explore a small but representative section of the energy landscape, from which we

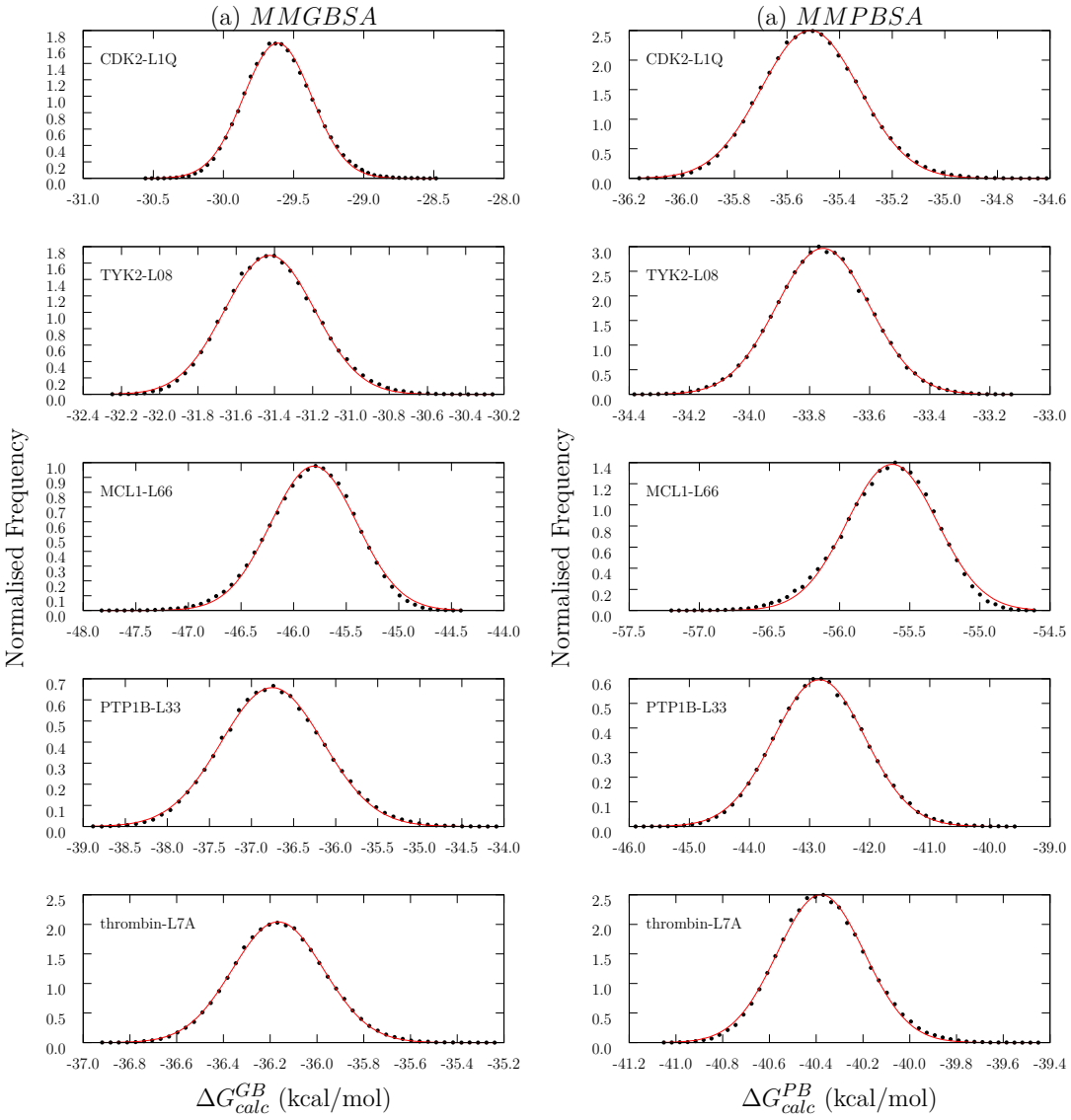


Figure 2.3: *Normalized frequency distributions of binding affinities obtained by ESMACS for 5 different receptor-ligand systems, which are specified in the top left corner of each graph: a) is the distributions for the MMGBSA and b) using the MMPBSA method. Each data point corresponds to one frame from which binding affinities were generated. The chaotic dynamics of receptor-ligand systems allow us to compute, from Gaussian distributions, reliable probabilistic binding free energies.*

are able to obtain accurate and/or precise thermodynamic properties. Ensemble averaging, through the application of the ergodic hypothesis, connects the macroscopic scale of experiment, to simulation, which is executed on the microscopic scale.

Obtaining thermodynamic properties, via ensemble averaging, is achieved by molecular simulation, which explores the energy landscape of the system of study. A variety of molecular simulation algorithms have been developed, but the two most common methods in biomolecular modelling are Monte Carlo (MC) and molecular dynamics (MD). A brief explanation of MC simulations will be provided, but much of this section will focus on MD, as this is generally the preferred choice in biomolecular modelling, and is a major component of the methodology applied in this thesis.

2.4.1 Monte Carlo

The basis of MC is random exploration of configurations of a system, where the sample state M is only dependent on the preceding state, and has no bearing on the succeeding state. Thus, an MC simulation is a stochastic and time-independent sampling approach, and cannot give information of the evolution of a dynamical system.

MC simulations generate random configurations using a set of criteria that either accepts, or rejects, the succeeding configuration. If the move is accepted, than the simulation proceeds, but if a move is rejected, then another iteration is performed where a different configuration is chosen at random. The probability of the succeeding configuration being selected is equal to the Boltzmann factor, $e^{-\mathcal{V}(\mathbf{r}^N)/kT}$, where $\mathcal{V}(\mathbf{r}^N)$ is the potential energy of a system. The accepted configurations are then used to calculate an ensemble average of the thermodynamic property in question, by averaging over the total sample states (M):

$$\langle G \rangle = \frac{1}{M} \sum_{i=1}^{M} G(\mathbf{r}^N) \tag{2.30}$$

When a new random configuration is chosen, the potential energy is computed ($\mathcal{V}(\mathbf{r}^N)$). If the succeeding move is accepted, then $\mathcal{V}(\mathbf{r}^N)$ of the succeeding state is lower than the energy of current configuration. If $\mathcal{V}(\mathbf{r}^N)$ is higher in the succeeding state, the the following occurs. The Boltzmann factor of the difference between the old and new move are calculated:

$$e^{-(\mathcal{V}_{new}(\mathbf{r}^N) - \mathcal{V}_{old}(\mathbf{r}^N)/kT} \tag{2.31}$$

where $\mathcal{V}_{new}$ and $\mathcal{V}_{old}$ is the potential energy of the new and old configuration, respectively. A random number, between 1 and 0 is also generated. If the random number is lower than the value obtained from Eqn. 2.31, then the move is accepted, otherwise, the move is rejected. The most common used MC algorithm is the Metropolis algorithm [47].

MC simulations are preferred for simulations of gases or other lower density systems. This is because the large energy barriers in such molecules (e.g. torsional rotations) are easier to overcome because of the ability to make random moves. MC can also be used in simulations which vary in the number of particles (Grand Canonical MC), by adding moves for the creation or annihilation of particles [48]. MC would not be suitable for simulating condensed phase systems, like ligand-receptor complexes, because there is a large probability that the random move would result in the overlap of molecules. This means that a large number of moves will be rejected which will decrease the sampling efficiency. MC is also not suitable for studying phenomena that are dependent on time, like transport properties.

2.4.2 Molecular dynamics

The relative mass of nuclei, compared to electrons, is used as a reasonable approximation where atoms can be modeled as classical particles, and classical Newtonian mechanics can be applied. MD applies this theory to give a time-dependent sampling algorithm, with a temporal relation between configurations.

Newton's laws of motion

MD generates configurations by integrating Newton's second law of motion, $\mathbf{F}_i = m_i\mathbf{a}_i$, where $\mathbf{F}_i$ is the force felt by an atom, and m and $\mathbf{r}$ are the mass and acceleration of that atom. The differential form is:

$$-\frac{\partial \mathcal{V}}{\partial \mathbf{r}_i} = m_i \frac{d^2\mathbf{r}_i}{dt^2} \tag{2.32}$$

where $\mathcal{V}$ is the potential energy of the atom at position $\mathbf{r}$, and $d^2\mathbf{r}_i/dt^2$ is the second derivative of the position ($\mathbf{r}$) of an atom, with respect to time, t, which equates to the acceleration of the atom, $\mathbf{a}_i$. Eqn. 2.32 also demonstrates the important relationship between force, $\mathbf{F}$, and the potential energy, $\mathcal{V}$: the force felt by an atom, is equivalent to the negative gradient of the potential energy, with respect to the position of that atom, $-\partial\mathcal{V}/\partial\mathbf{r}_i$.

Updating atomic coordinates

Finite difference methods are used to integrate Newton's second law, and thus obtain a time-dependent trajectory of configurations. The MD integral is solved numerically, by breaking the integral into small frames, from which positions, velocities and accelerations of all atoms in a configuration are recorded. The fixed time period between the frames is known as the time step, δt. The integration process can be broken down like so:

- the force of each atom in a configuration is calculated at time t
- the accelerations of all atoms are determined at time t
- from this, the positions ($\mathbf{r}$) and velocities ($\mathbf{v}$) are calculated, also at time t
- $\mathbf{r}(t)$ and $\mathbf{v}(t)$ are then used to calculate the $\mathbf{r}(t+\delta t)$ and $\mathbf{v}(t+\delta t)$
- the forces of each atom at time $t+\delta t$ are computed

which is iterated for a given time period. There are many finite difference algorithms [49, 50, 51] used to generate trajectories, and all are approximated as a Taylor series expansion. Here, we will describe the velocity Verlet algorithm [52] (Eqn. 2.33) which is thought to be the best MD algorithm, as it is precise, and generates $\mathbf{r}$, $\mathbf{v}$ and $\mathbf{a}$ in the same time step.

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 \tag{2.33a}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \frac{1}{2}[\mathbf{a}(t) + \mathbf{a}(t+\delta t)]\delta t \tag{2.33b}$$

Eqn. 2.33b states that the acceleration must be known at time t, and $t+\delta t$, and so the velocity Verlet algorithm is performed in three steps.

$$\mathbf{v}(t+\frac{1}{2}\delta t) = \mathbf{v}(t) + \frac{1}{2}\delta t\mathbf{a}(t) \tag{2.34a}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t+\frac{1}{2}\delta t) + \frac{1}{2}\delta t\mathbf{a}(t+\delta t) \tag{2.34b}$$

First, the positions of all atoms, at time $t+\delta t$ are determined using Eqn. 2.33a, using the velocities at accelerations at time t. Secondly, velocities at time $t+\frac{1}{2}\delta t$ are calculated (Eqn. 2.34a). The forces can be calculated from the current time step, to give the acceleration of the atoms at time $t+\delta t$, and finally the velocities at $t+\delta t$ are determined using Eqn. 2.34b.

MD at constant temperature and pressure

A standard MD simulations is performed using an NVE ensemble, where the number of atoms (N), the volume (V) and energy (E) remain constant, but the temperature and pressure are allowed to change. In the context of the protein-ligand binding process, the experimental environment should be replicated, and here the temperature (T) and pressure (P) are constant, and the energy fluctuates. This introduces the NPT, or isothermal-isobaric ensemble, where the atomic positions of a simulation are updated on the basis of constant temperature and pressure. In MD simulations, the average temperature and pressure of configurations are kept constant using thermostats and barostats, respectively.

The total energy of a system is the sum of the kinetic and potential energies, which can be calculated from the positions and velocities of all atoms:

$$E_{tot} = \sum_{i=1}^{N} m_i \mathbf{v}_i^2 + \mathcal{V}(\mathbf{r}_i) \tag{2.35}$$

where the temperature of a system is proportional to the average kinetic energy, $\langle E_{kin} \rangle = 3kT/2$. Here, $\langle E_{kin} \rangle$ is the average kinetic energy. Since, in NPT ensembles, the average energy of the system is kept constant (Eqn. 2.35), and the the potential energy is dependent on the positions of the atoms, the kinetic energy, and thus the temperature, will fluctuate considerably to compensate.

The Berendsen thermostat [53] is a way of maintaining the average temperature of a system, where the simulation unit cell is coupled to a heat bath. The velocities of the atoms are scaled so that the change in temperature of the system is equal to the difference in temperature between the simulation unit cell and the heat bath:

$$\frac{dT(t)}{dt} = \frac{1}{\tau_T}(T_{bath} = T_{actual}(t)) \tag{2.36}$$

where τ_T is the coupling parameter that determines the magnitude of the coupling between unit cell and heat bath, T_{bath} is the temperature of the heat bath, and $T_{actual}(t)$ is the temperature of the unit cell at the current time step.

$$\lambda_T^2 = 1 + \frac{dt}{\tau_T}\left(\frac{T_{bath}}{T_{actual}(t)} - 1\right) \tag{2.37}$$

The scaling factor, shown in Eqn. 2.37, is λ_T^2. When τ_T is large then the scaling factor is weak, and vice versa.

Scaling the velocities does not overcome the phenomenon of "hot solvent, cold solute" where the thermal energy is not distributed evenly throughout the system, resulting in different temperatures for the solvent and solute. Langevin dynamics, explained later in the next section, resolves this problem.

Maintaining constant average pressure of a system is tackled in a similar way to a Berendsen thermostat. The Berendsen barostat [53] scales the volume of the system, by adjusting the coordinates of the atoms. This has the same form as the thermostat where the change in pressure of the system is equal to the difference in pressure between the simulation unit cell and the 'pressure bath'.

$$\frac{dP(t)}{dt} = \frac{1}{\tau_P}(P_{bath} = P_{actual}(t)) \tag{2.38}$$

Here, τ_P is the coupling constant and P_{bath} and P_{actual} are the pressures of the 'bath' and unit cell, respectively.

$$\lambda_P = 1 - \kappa\frac{dt}{\tau_P}(P_{bath} - P_{actual}(t)) \tag{2.39a}$$

$$\mathbf{r}_i' = \lambda_P^{1/3}\mathbf{r}_i' \tag{2.39b}$$

The volume of the system is scaled by λ_P (Eqn. 2.39a), and the compressibility factor κ equals $-1/V(dV/dP)_T$. The updated atomic positions are given by Eqn. 2.39b.

Langevin dynamics

The benefit of Langevin dynamics (LD) is that primary focus can be turned away from the solvent, and onto the solute. LD allows to incorporate the effects of solvent without the requirement of explicit solvent molecules to be present. Thus a simulation using LD means that solvent effects can be taken into account, which also influences the dynamics of the solute. This happens through random collisions and a frictional drag force places on the solute, as it moves through the solvent.

As cited earlier, LD can be used to overcome "hot solvent, cold solute" effects. By introducing a Langevin thermostat [54, 55], thermal energy from the heat bath, can be transferred to the unit cell through collisions between the atoms in the heath bath, and that of the unit cell. LD, shown in Eqn. 2.40, has three components that contribute to the force exerted on an atom. The first component on the right side of Eqn. 2.40, is the force exerted by other atoms, and is described by the potential energy. The second, is the force of an atom moving through solvent, which is modeled by a frictional drag due to the solvent, where ζ is the friction coefficient. The last term is the force is due to random fluctuations in the interactions with solvent.

$$m\frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i(\mathbf{r}_i(t)) - \zeta_i\frac{d\mathbf{r}_i(t)}{dt}m_i + \mathbf{R}_i(t) \tag{2.40}$$

The random force adds thermal energy into the system, and is thus associated with temperature, and the frictional force removes thermal energy from the system. There are some assumptions made in the LD model. First, it is assumed that the frictional coefficient, ζ, has no bearing on the time and position of the atoms.

Second, the random force is independent of the velocities of the particles and is taken to have a Gaussian distribution with zero mean. Due to the random force component, implementing a Langevin thermostat in an MD algorithm, means that the simulation is no longer deterministic.

Periodic boundary conditions

In simulations methods, an appropriate definition of a boundary enables the correct determination of macroscopic properties with the use of a relatively small number atoms. Periodic boundary conditions [40, 39] makes this possible in such a way that the forces exerted on the atoms are the same as if they were in bulk fluid. Without this condition, a significantly larger number (technically an infinite number) of atoms would be required to achieve a solution to macroscopic properties.

The basic principle of periodic boundary conditions is that if a solvent molecules leaves the simulation unit cell, which is most commonly cubic in shape, from one side, it will re-enter from the opposing side. The purpose of this is that the system should not feel the effects of the boundaries. This is achieved by replicating the cubic box of the simulation unit cell, in all directions. The coordinate of the atoms in the imaginary unit squares can be calculated by adding or subtracting the integral multiples of the simulation box length. The box size needs to be large enough so that the protein atoms in the simulation unit cell, do not feel an effect from the protein atom of any periodic image. A key consideration in periodic systems is the treatment of long and short range forces, which will be described next.

Non-bonded cut-offs

The potential energy of a system is the sum of the bonded and non-bonded interactions ($\mathcal{V}_{total} = \mathcal{V}_{bonded} + \mathcal{V}_{non-bonded}$). The number of bonded interactions that need to be calculated are proportional to the number of atoms ($3N - 6$), whereas

non-bonded terms increase in the order of N^2. Therefore, the most laborious term to calculate is the non-bonded contributions to the potential energy. In theory, non-bonded interactions should be generated for all pairs of atoms, but this is not always required. A way to truncate the potential is by: (a) introducing a non-bonded cut-off where the potential is set to zero for any interaction past the cut-off distance, and (b) apply the minimum image convention, where each atom interacts with only one image of itself, within the cut-off distance (this is repeated infinitely using periodic boundary conditions).

Cut-offs alone do not improve the efficiency of computing the potentials, because this would still require the distances between each pair of atoms to be determined. In liquid phase simulation, the atoms within a cut-off do not fluctuate much over 10-20 time steps, meaning that distances between atoms are calculated less frequently, and thus improving the compute efficiency. The non-bonded neighbour list stores the information of all atoms within the cut-off distance, and additional atoms that are slightly over the cut-off.

An important consideration is the frequency at which the neighbour list is updated. If it is updated too frequently, then this compromises the efficiency, and if the neighbour list is not updated enough, then incorrect energies will be calculated. Further, setting the potential to zero so suddenly, introduces a discontinuous energy potential. This can be overcome by using a switch function [56] which gradually tapers off the potential energy close to the cut-off distance.

The long-range electrostatic term requires the most compute time as the r^{-1} term decays slowly with respect to distance. Further, when periodic boundary conditions are applied, cut-offs cannot be so large that an atom interacts with its own image, and so this limits cut-offs to less than half of the simulation unit cell. Slow decay of electrostatic interactions mean that considerable contributions are made to the potential at distances greater than half of the unit box length.

The Ewald summation [57] exploits the periodicity of the simulation unit cell, and splits the potential into near- and far-field contributions. The near-field contribution is obtained by taking a Gaussian function centered at each atom, with an opposing point charge. This has a screening effect on the atomic charges allowing for rapid convergence of near-field contributions. To reinstate the original point charge interaction, the screening potentials are subtracted again. This compensating term is an interaction between Gaussian distributions, and is thus a far-field interaction. This can be efficiently evaluated as the sum of the Fourier transforms of the potential and charge density. The Ewald Summation reduces the scaling from N^2 to $N^{3/2}$. A related approach, the Particle Mesh Ewald method [58] further improves scaling to $N \ln(N)$.

Constrained dynamics

The limitation, in terms of time step selection, is the speed of the fastest processes in a simulation: bond vibrations. A time step of 1 femtosecond (10^{-6} s) is required to capture the bond vibrations in a simulation. Recording atomic information at this frequency, is computationally inefficient and would require a vast amount of time steps to gain accurate representations of biological phenomena.

The stretching vibration of hydrogen atoms bonded to heavy atoms are the fastest of the degrees of freedom, but has a relatively small influence on the energy outputs. By treating the bonds between heavy and hydrogen atoms rigid, the time step can be increased to 2 or 3 femtoseconds. Although this seems a small change, over the course of a simulation this contributes considerably and enables longer, or more simulation replicates, to be performed at the same computational expense.

The SHAKE [59, 60] and RATTLE [59] algorithm are common constraint methods incorporated into MD algorithms. Here, the positions of all atoms are determined using Newton's equation of motion, with no constraints. The atomic positions are then corrected by the method of Lagrange determined multipliers. Another

method, SETTLE [61], has specifically been developed to constrain bonds in solvent molecules.

2.5 Molecular mechanics force fields

An MD algorithm computes the force experienced by each atom. To compute the force, it is required to know the potential energy of each atom. A molecular mechanics force field is a potential energy function that defines the intramolecular forces using classical mechanics. The following is a breakdown of a force field into its component parts.

2.5.1 A force field model

A force field can be segmented as follows:

$$\mathcal{V}_{total} = \mathcal{V}_{bonded} + \mathcal{V}_{non-bonded} \tag{2.41a}$$

$$\mathcal{V}_{bonded} = \mathcal{V}_{stretch} + \mathcal{V}_{angle} + \mathcal{V}_{dihedral} \tag{2.41b}$$

$$\mathcal{V}_{non-bonded} = \mathcal{V}_{vdW} + \mathcal{V}_{elec} \tag{2.41c}$$

where the total potential energy $\mathcal{V}_{total}$ is the sum of the bonded ($\mathcal{V}_{bonded}$) and non-bonded ($\mathcal{V}_{non-bonded}$) intramolecular forces. The bonded forces can be broken down even further into bond stretching ($\mathcal{V}_{stretch}$), bending ($\mathcal{V}_{angle}$) and torsional ($\mathcal{V}_{dihedral}$) terms. The bonded terms are represented as a deviation of the bond length (r), angle (θ) or dihedral (γ) from an equilibrium value.

$$\mathcal{V}_{stretch} = \sum_{\text{stretch}} K_r (r - r_{eq})^2 \tag{2.42a}$$

$$\mathcal{V}_{angle} = \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 \tag{2.42b}$$

$$\mathcal{V}_{dihedral} = \sum_{\text{dihedral}} \frac{V_n}{2}[1+\cos(n\phi-\gamma)] \tag{2.42c}$$

The $\mathcal{V}_{stretch}$ term is modeled by the harmonic potential, that is Hooke's law formula, where r_{eq} is the equilibrium bond length, and K_r is the stretching constant. Another expression that derives realistic energy profiles is the Morse potential, however this is used in systems where bond distances deviate significantly from the equilibrium values, which is not often seen in biological systems, therefore, Hooke's law usually suffices. $\mathcal{V}_{angle}$ describes the deviations of bond angles from their reference values, and is also commonly described by Hooke's law, where θ_{eq} is the equilibrium bond angle, and K_θ is the bending constant. It requires much more energy to stretch bonds, than it does to bend them, so force constants will be much larger in the first term. $\mathcal{V}_{dihedral}$ expresses the torsional angles, that is, the rotation around the $B-C$ bond of a molecule, $A-B-C-D$. Torsional or dihedral potentials are usually defined using a cosine series expansion, where ϕ is the dihedral angle, n is the multiplicity which defines the number minimum points when the bond is rotated 360°, γ shows where the dihedral angle passes through a minima, and V_n is the barrier height.

Similarly, the non-bonded interactions can be broken down in to a short-range van der Waals term ($\mathcal{V}_{vdW}$), and a long-range electrostatic ($\mathcal{V}_{elec}$) term:

$$\mathcal{V}_{vdW} = \sum_{i<j}\left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right] \tag{2.43a}$$

$$\mathcal{V}_{elec} = \sum_{i<j}\frac{q_i q_j}{\epsilon r_{ij}} \tag{2.43b}$$

The van der Waals force is described by the 12-6 Lennard-Jones equation which models the attractive and repulsive interactions depending on the inter-atomic separation, $\mathbf{r}_{ij}$. As $\mathbf{r}_{ij}$ decreases from infinity, the negative $1/r^6$ term dominates the interaction and so the atoms feel an attractive force as the energy becomes

more negative. As $\mathbf{r}_{ij}$ continues to decrease, and tends towards zero, the $1/r^{12}$ dominates and leads to an increase in energy, and thus repulsion between the two atoms. Coloumb's law is most commonly used to describe electrostatic interactions in macromolecular systems, where q_i and q_j are the atomic charges, ϵ is the relative permittivity of space and $\mathbf{r}_{ij}$ is the interatomic distance. A more detailed account on the implications of non-bonded interactions in molecular simulations, can be found in the section 2.4.2 under "Non-bonded cut-offs".

In addition to the functional form that is selected to model respective interactions, parameters are also included which describe geometric and energetic properties of the interaction [62]. These can be derived either using computational means (i.e. quantum mechanical calculations) or empirically. The combination of a functional form, and a set of parameters is a force field.

There are several molecular mechanics force fields that have been developed; all of these are based on a similar functional form to describe molecular interactions. The most widely used force fields are AMBER [63], CHARMM [64], GROMOS [65] and OPLS [66]. There are subtle differences between these force fields. For example, in the case of defining the energy of improper dihedral angles, OPLS and AMBER incorporate the improper dihedral term within the dihedral term seen in Eqn. 2.42c. CHARMM and GROMOS have an additional functional term that defines this molecular property.

Choosing a potential energy function for a protein is relatively simple compared with that of small drug-like molecules. The function for a protein is limited to the 20 amino acids that are commonly found in proteins, and the atom type and intramolecular forces that define them. These have been described and tested extensively, and so a realistic simulation of protein dynamics is available. Small molecules, however, have many more permutations with respect to atom types and degrees of freedom, and so careful parameterisation is required.

All of the aforementioned force field packages have a extensive range of defined atom types that are commonly seen in drug-like molecules. AMBER has the benefit, for end-users, that it is able to generate a force field model automatically and is compatible with the AMBER protein force field, whereas OPLS requires manual selection of atom type selection. Transferrability of parameters across a wide range of small molecules has been achieved and general force fields have been developed, such as General Amber Force Field (GAFF, [67]) and CHARMM General Force Field (CGenFF, [68]) which aim to incorporate the parmaterised small molecule within the respective protein force field. OPLS 2.1 is a force field, which has been applied to several biological systems [45] that relate to calculating binding free energies.

Applying accurate force field parameters to a biological system is crucial when preparing a MD simulation. The difference between the above force fields is subtle, however one has to pick carefully which force field is best suited dependent on the type of simulation.

2.5.2 Sources of error

The number of configurations sampled, then, is irrelevant if the model describing the intra and intermolecular force is not an accurate representation. This applies to both the protein and small-molecule that is simulated. Perhaps, a greater challenge is determining the correct parameters for small molecules that exhibit unconventional stereochemistry, or contain chemical groups, that are not extensively studied. Thus, the quality of the simulation is directly dependent on the accuracy of the molecular mechanics force field.

2.6 Free energy methods

As was described in section 2.3.3, it is challenging to determine entropic thermodynamic properties, such as the change in Gibbs free energy, ΔG, but it is possible to

calculate the difference of such properties. Hence, in the context of protein-ligand binding, we can calculate the change in ΔG which gives $\Delta\Delta G$. This is a relative binding affinity, as it is the calculation of the relative change in ΔG, between two systems. Techniques which employ this strategy, namely Thermodynamic Integration (TI, [69, 70]) and Free Energy Perturbation (FEP, [71]), are commonly known as 'exact' methods, and are computationally rigourous. An alternative, and computationally more efficient strategy, is to compute the absolute binding affinity (ΔG) through the application of a less accurate physical model, and empirically derived parameters. The most widely used of these approaches is Molecular Mechanics and the Poisson-Boltzmann Surface Area Approximation (MMPBSA, [72]). This yields binding affinities that are 'approximate', but exhibit precision. This section will detail the key 'exact' and 'approximate' free energy methods that are commonly applied in protein-ligand binding.

2.6.1 Exact methods

TI and FEP are related methods inasmuch that they both compute the relative binding affinity between two states, through a series of alchemical transformations each defined by a coupling parameter, λ. $\lambda = 0$ is the initial state and $\lambda = 1$ is the final state. The transformations are termed 'alchemical' because the intermediate states are unphysical, and would not be seen in experiment. However, as G is a state function, the path taken between the first and last λ state is irrelevant. Through the use of a thermodynamic cycle, relative changes in ΔG can be determined ($\Delta\Delta G$). Both methods require MD simulations for each λ window which is the reason for the high computational cost. They also require small increments of the intermediate λ state, because in FEP, an overlap in phase space is needed; and in TI, this allows for accurate numerical integration.

A property of a state function, such as G, is that regardless of the path taken from initial to final state, the sum of all paths required to close a thermodynamic cycle,

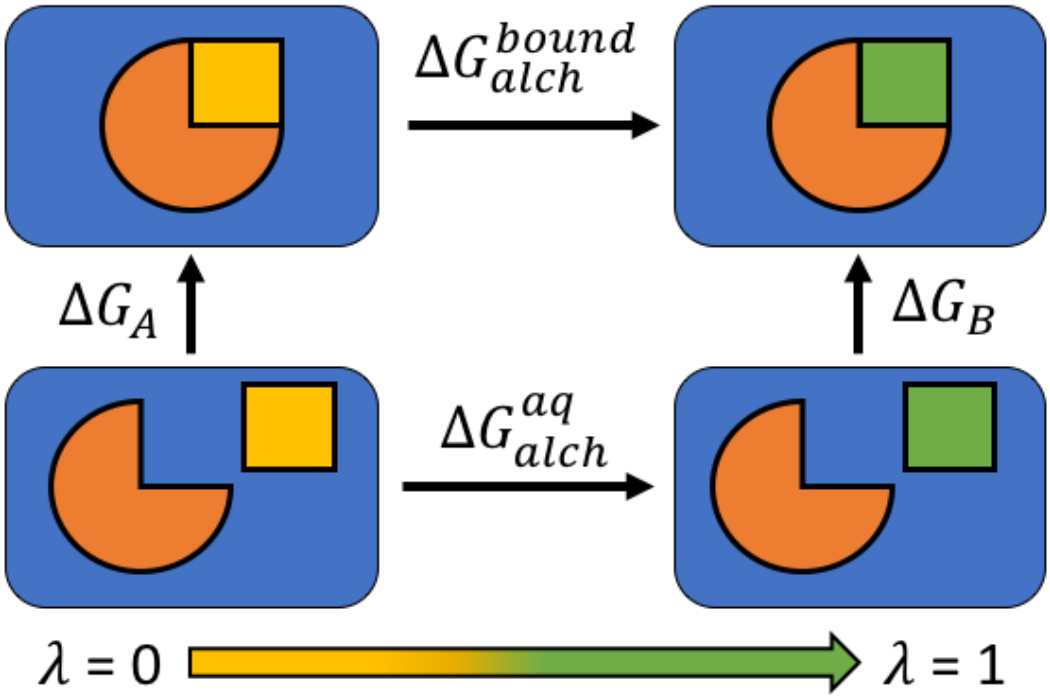


Figure 2.4: *A representation of the the thermodynamic cycle used in the computation of relative binding affinities ($\Delta\Delta G_{bind}$) using exact methods, namely TI and FEP.*

will always equal to zero. Exact methods employ the thermodynamic cycle in Fig. 2.4, to calculate $\Delta\Delta G_{bind}$. A logical approach is to compute the difference of the processes that represent ΔG_A and ΔG_B:

$$\Delta\Delta G_{bind} = \Delta G_B - \Delta G_A \tag{2.44}$$

where ΔG_B and ΔG_A are the processes of ligand A and ligand B, binding to a protein. Simulating this process is cumbersome because the path involves large conformational rearrangements of the ligand binding to the protein, plus the conformational changes of the protein associated with the binding process. Modeling this process would require an extremely long simulation to obtain converged ΔG values.

A counter-intuitive approach is to simulate the alchemical transformation of ligand A, to ligand B, bound to the protein, ΔG^{bound}_{alch}, and the transformation of the two entities dissociated in aqueous solvent, ΔG^{aq}_{alch}. Since the ΔG value of the protein will not change considerably, as it is the same in the initial and final state, only a simulation of the alchemical transformation of the ligand is required. This requires the transformations of chemically related ligands, so that conformation changes are small, and converged ΔG values are obtained.

$$\Delta\Delta G_{bind} = \Delta G^{aq}_{alch} - \Delta G^{bound}_{alch} \tag{2.45}$$

As such, using the thermodynamic cycle, we can calculate the relative binding affinity, of two chemically similar ligands, using alchemical free energies. This is formulated in the above equation.

2.6.1.1 Free energy perturbation

FEP computes $\Delta\Delta G$ by averaging over finite differences of λ in the potential energy function. The difference in free energy between two states is defined like so:

$$\Delta G = \Delta G_A - \Delta G_B = -kT \ln\langle e^{-(\mathcal{V}_B - \mathcal{V}_A)/kT}\rangle_M \tag{2.46}$$

Since the alchemical transformation from final to end state is broken down into λ states, the potential energy is computed by summing the change in each λ window:

$$V(\lambda, \mathbf{r}) = (1-\lambda)V_A(\lambda, \mathbf{r}) + \lambda V_B(\lambda, \mathbf{r}) \tag{2.47}$$

where the potential energy of ligand A and B is defined as $V_A(\lambda, \mathbf{r})$ and $V_B(\lambda, \mathbf{r})$, and $V(\lambda, \mathbf{r})$ is the potential energy of the intermediate state.

Recently, there has been an advance in FEP-based methods that has led to considerable interest [45]. FEP+ uses FEP and replica exchange solute tempering (REST, [46]), where the hamiltonian between different λ windows are exchanged to enhance the sampling of the simulation, particularly in the region of ligand binding. However the predictions that are made are based on a single simulation for each ligand transformation. Further, the error is calculated by averaging the error around a thermodynamic cycle, involving numerous perturbations. This method spreads error evenly of a number of perturbations, and yields an unrepresentative error for a particular FEP calculation.

2.6.1.2 Thermodynamic integration

TI calculates the relative binding affinity by averaging over a differentiated energy function, with respect to the intermediate λ states.

To generate the free energy changes of an alchemical transformation, the derivative of the total potential energy with respect to λ is computed, and then integrated numerically for all lambda states (Eqn. 2.48). This is done for both the ligand in aqueous solution, ΔG^{aq}_{alch}, and the ligand-protein complex, ΔG^{bound}_{alch}:

$$\Delta G_{alch} = \int_0^1 \left\langle \frac{\partial V(\lambda, \mathbf{r})}{\partial \lambda} \right\rangle d\lambda \tag{2.48}$$

Then, the thermodynamic cycle is used (Fig. 2.4) to determine the relative binding affinity, $\Delta\Delta G_{bind}$, between ligand A and B, which is equal to the the difference in ΔG between the free (ΔG^{aq}_{alch}), and bound ΔG^{bound}_{alch} ligand (Eqn. 2.45).

A recently published method, "Thermodynamic Integration with Enhanced Sampling" (TIES, [42]), runs replica simulations for each λ window of each ligand transformation, which allows for tighter control over standard errors and thus produces more reliable results. Running replica simulations means that the error is representative of the perturbation in question.

2.6.1.3 Absolute binding affinity predictions

It is also possible to create a thermodynamic cycle that enables the determination of absolute binding affinities. Following a similar approach described in Fig. 2.4, calculating absolute binding affinities requires only two simulations, that is the disappearance of a ligand in aqueous solution and the disappearance of a ligand within a complex. The initial and final λ states would then be the ligand, and absence of ligand, respectively.

Historically, absolute binding affinities have not been shown the same levels of interest as relative binding affinity predictions, but a recent study conducted by Aldeghi and colleagues [44] performed FEP-based absolute binding free energy calculations on a diverse range of ligands. Good correlation were reported with experimental binding affinities. However, this approach was tested on a rigid drug target system, resulting in fast converging properties. Tight error bars were reported, albeit from a single simulation, which is due to the rigidity of the protein target. Recently, an ensemble-based method to determine absolute binding affinities has also been developed [73].

2.6.2 Approximate methods

The computational cost of executing the so-called 'exact' methods has led to alternative, more computationally efficient approaches, involving empirical parameters, and some general assumptions. In the context of drug discovery programmes, the requirement to predict binding affinities for a large library of drugs has led to the development of 'approximate' methods, which is thought to be a good balance between computational cost and accuracy. However, there is an inverse relationship between computational efficiency, and accuracy which needs to be considered. The following 'approximate' methods that are explored, are presented in the order of increasing computational cost, and thus increasing accuracy.

2.6.2.1 Molecular docking and scoring functions

The purpose of molecular docking and subsequent free energy scoring functions, are to predict the binding affinity of large libraries of compounds to a protein target [74]. This is more commonly known as virtual screening and has been employed in drug discovery as a strategy to rapidly assess binding affinities. Binding affinity assessment on this scale brings two main problems. The 'docking problem' is where realistic generation and evaluation of structures is difficult without enhanced sampling methods. Secondly, generating a realistic scoring function requires the evaluation of many energy terms, but this is not possible in virtual screening due to the large computational time required to calculate all energy terms. For these reasons, virtual screening techniques are highly approximate, and are, at best, suitable as a first approximation of binding affinities before more rigorous techniques are used [75].

To achieve efficient computation, docking algorithms usually employ a rigid protein structure and allow the ligand to explore conformations. The obvious issue here is that proteins often experience significant conformational changes upon ligand binding, which is ignored. With regard to scoring functions, the computational requirements in calculating all energy terms involved in binding means only some energetic terms are considered. For example, the internal bonded interactions described by molecular mechanics is often used, but the non-bonded terms are usually predicted by linear dependence to polar and non-polar regions.

There are a number of molecular docking and scoring functions available with varying performance [76, 77, 78, 79, 80, 81].

2.6.2.2 Linear activation energy

The linear interaction energy (LIE, [82]), also know as the linear response (LR) method is a semi-empirical approach that requires sampling of only two states: the

protein-ligand complex in solvent, and the free ligand in solvent. LIE takes into account only the electrostatic and van der Waals potential interactions between the solute and solvent.

$$\Delta G_{bind}^{LIE} = \alpha \Delta \langle \mathcal{V}_{elec} \rangle + \beta \Delta \langle \mathcal{V}_{vdW} \rangle + \gamma \tag{2.49}$$

The free energy is obtained by linearly fitting the the difference in $\mathcal{V}_{elec}$ and $\mathcal{V}_{vdW}$ between the ligand and its environment, either protein or solvent. The parameters α, β and γ are empirically derived, depending on the protein and ligand that is under study. The reliance of experimentally derived parameters for new ligands means that this method is a poor strategy for drug discovery purposes.

2.6.2.3 Molecular Mechanics and the Poisson-Boltzmann Surface Area approximation

Molecular Mechanics and the Poisson-Boltzmann Surface Area approximation (MMPBSA) is the most computationally rigorous of the approximate methods described. MMPBSA has been a popular strategy for predicting absolute binding affinities for protein-ligand systems. This is due to the relatively inexpensive computational requirements, and good comparisons in binding affinities between different ligands bound to the same protein. A minimum requirement of MMPBSA is a single simulation of the protein-ligand complex, compared with TI and FEP, which require simulations for each λ state.

MMPBSA computes the average binding free energies of the end-states in a protein-ligand system. These being the initial state, the ligand and protein unbound in aqueous solvent; and final state, the bound protein-ligand complex. These free energies are then used to determine the ΔG value:

$$\Delta G_{bind} = \langle G_{complex} \rangle - \langle G_{protein} \rangle - \langle G_{ligand} \rangle \tag{2.50}$$

where $\langle \cdots \rangle$ denotes ensemble averages of the three entities, from a representative sample of configurations obtained from an MD simulation.

The purpose of the MMPBSA approach is to compute the free energy of the association of a ligand and protein, in aqueous solution, ΔG^{aq}_{bind}. The problem that arises here, is that the majority of the energetic contributions will come from the solvent-solvent interactions, giving rise to large fluctuations in ΔG^{aq}_{bind}. To achieve converged results, an inordinate number of configurations need to be sampled.

To overcome this, a thermodynamic cycle (Fig. 2.5) can be designed to circumvent the simulation of the ΔG^{aq}_{bind} process. Instead, the binding free energy *in vacuo*, ΔG^{vac}_{bind}, is computed, in addition to the free energies of solvation of the complex (ΔG^{sol}_{com}), protein (ΔG^{sol}_{rec}) and ligand (ΔG^{sol}_{lig}).

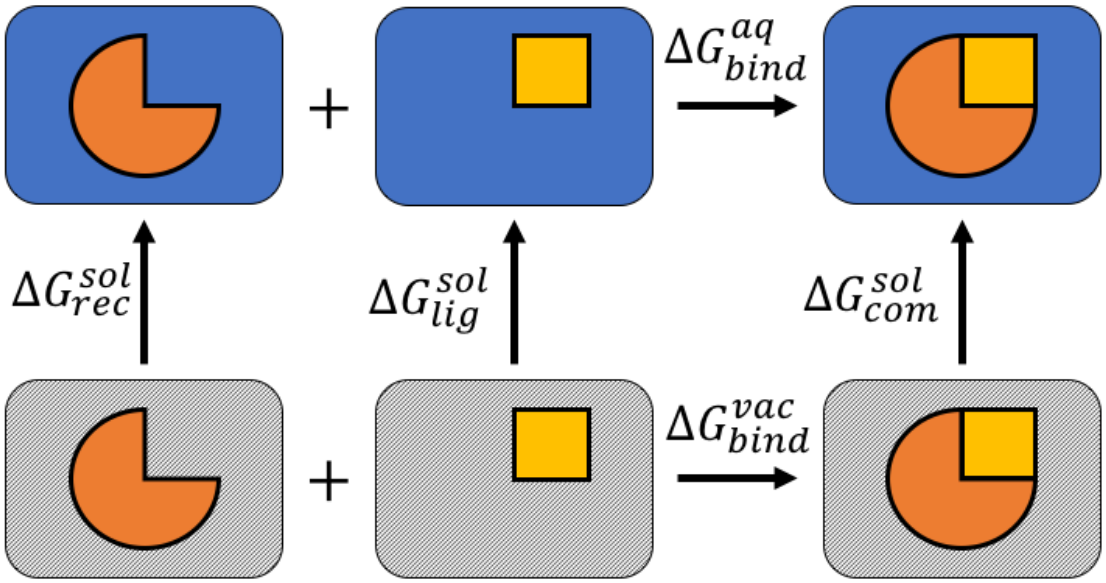


Figure 2.5: *A representation of the the thermodynamic cycle used in the computation of absolute binding affinities (ΔG_{bind}) using the approximate method, MMPBSA.*

As a result, using the thermodynamic cycle in Fig. 2.5, ΔG^{aq}_{bind} can be calculated like so:

$$\Delta G^{aq}_{bind} = \Delta G^{vac}_{bind} + \left(\Delta G^{sol}_{com} - \Delta G^{sol}_{rec} - \Delta G^{sol}_{lig}\right) = \Delta G^{vac}_{bind} + \Delta G^{sol} \tag{2.51}$$

The binding free energy *in vacuo* is the summation of the electrostatic, (ΔG^{MM}_{elec}), van der Waals (ΔG^{MM}_{vdW}) and intramolecular interactions (ΔG^{MM}_{int}):

$$\Delta G^{vac}_{bind} = \Delta G^{MM}_{int} + \Delta G^{MM}_{vdW} + \Delta G^{MM}_{elec} \tag{2.52}$$

and is calculated using the molecular mechanics force field described in section 2.5.

$$\Delta G^{sol} = \Delta G^{sol}_{non-pol} + \Delta G^{sol}_{pol} \tag{2.53}$$

The free energy of solvation term, ΔG^{sol}, is the free energy of transferring the protein, ligand or complex, from a vacuum, into an aqueous solvent. This term can be broken down even further to a polar (electrostatic) and non-polar (van der Waals) contributions (Eqn. 2.53). The polar contribution is calculated by treating the solute as a continuous region of low dielectric constant, and the solvent is represented as a constant high dielectric continuum, enabling numerical solution of the linear Poisson-Boltzmann equation. The non-polar term is calculated from empirical parameters related to the solvent accessible surface area.

Linear Poisson-Boltzmann equation

The Poisson-Boltzmann (PB, [72]) equation is used to determine ΔG^{sol}_{pol}, by combining the Poisson equation, which relates charge density, ρ, to the electrostatic potential of a changing dielectric medium, and the Boltzmann distribution for mobile ions.

The electrostatic potential, in a medium of changing dielectric value with respect to position, can be related to the charge density:

$$\nabla \cdot [\epsilon(\mathbf{r}) \nabla \phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}) \tag{2.54}$$

The above Poisson equation (Eqn. 2.54) describes the local electrostatic potential, $\phi(\mathbf{r})$, generated by the charge density at that localised region, $\rho(\mathbf{r})$, where the charge density, ρ, is the distribution of charge across the system. In biomolecular systems, the variation in dielectric constant $\epsilon(\mathbf{r})$ will be between the solute, which is generally between 1 and 4, and the solvent, usually set at approximately 80.

The Poisson equation requires modification to incorporate the ionic distribution in solution, as a response to the electrostatic potential. Thus, negative ions will accumulate where the potential is positive, and vice versa. Over accumulation of ions is compensated due to natural thermal motion which is represented as a Boltzmann distribution:

$$\rho_{\pm} = \pm qce^{-q\phi/kT} \tag{2.55}$$

where $\rho_{\pm}$ is the charge density, $\pm q$ is the ionic charge and c is the concentration. Combining Eqn 2.54 and 2.55, gives the Poisson-Boltzmann equation:

$$\nabla \cdot [\epsilon(\mathbf{r}) \nabla \phi(\mathbf{r})] - \kappa'^2 sinh[\phi(\mathbf{r})] = 4\pi\rho(\mathbf{r}) \tag{2.56a}$$

$$\kappa^2 = \frac{\kappa'^2}{\epsilon} = \frac{8\pi N_A e^2 I}{1000\epsilon kT} \tag{2.56b}$$

where κ'^2 is related to the Debye-Hückel equation, which takes into account the interaction energy of ions: e is the electronic charge, I is the ionic strength of the solution, and N_A is Avogadro's number. This equation can be linearised by

expanding the sine function as a Taylor series expansion (Eqn. 2.56), and selecting only the first term:

$$\nabla \cdot [\epsilon(\mathbf{r}) \nabla \phi(\mathbf{r})] - \kappa'^2 \phi(\mathbf{r}) \left[1 + \frac{\phi(\mathbf{r})^2}{6} + \frac{\phi(\mathbf{r})^4}{120} + \cdots \right] = 4\pi\rho(\mathbf{r}) \tag{2.57a}$$

$$\nabla \cdot [\epsilon(\mathbf{r}) \nabla \phi(\mathbf{r})] - \kappa'^2 \phi(\mathbf{r}) = 4\pi\rho(\mathbf{r}) \tag{2.57b}$$

consequently, leading to the linearised Poisson-Boltzmann equation.

The derivatives in Eqn. 2.57b are determined by solving a partial differential equation for the potential, $\phi(\mathbf{r})$. A cubic lattice is superimposed on the solute and solvent, and values for the dielectric constant, charge density, electrostatic potential, and ionic strength are assigned to each grid point (i.e. the centre of each grid cube). The atomic charge does not always fall on a grid point, and so the charge is distributed to 8 surrounding grid points. As the atomic charge nears a grid point, a greater proportion of the total charge of that grid is assigned to that atomic charge. The boundary between solute and solute are defined as the molecular surface, and solvent accessible surface. Grid points that fall within the solute are assigned a dielectric constant representative of the solute, and grid points that fall outside this are assigned a high dielectric constant, representative of water.

The ΔG_{pol}^{sol} term is calculating by performing two calculations, using the same grid points and solute dielectric, but different exterior dielectric constants; water and vacuum:

$$\Delta G_{pol}^{sol} = \frac{1}{2} \sum_i q_i \left(\phi_i^{wat} - \phi_i^{vac} \right) \tag{2.58}$$

where q_i is the charge assigned to each point on the cubic grid, and ϕ_i^{wat} and ϕ_i^{vac} are the electrostatic potentials in water and vacuum respectively, at the same point. This calculation is performed for all three entities in the protein-ligand

binding process, that is the protein, ligand and complex.

Generalised Born equation

The Born equation is an alternative way of calculating the electrostatic contribution to the free energy of solvation using continuous dielectric media:

$$\Delta G_{elec}(q) = -\left(1 - \frac{1}{\epsilon}\right)\frac{q^2}{2a} \tag{2.59}$$

where q is the net charge and a is the radius of the cavity. The Born equation treats positive and negative ions the same which is not representative of biomolecular solutes in solvent, but a variation of Eqn. 2.59, where partial atomic charges are incorporated leads to the generalised Born (GB, [83]) equation:

$$G_{elec}(q_i, q_j) = -\left(1 - \frac{1}{\epsilon}\right)\frac{q_i q_j}{f_{ij}} \tag{2.60a}$$

$$f_{ij} = \sqrt{\mathbf{r}_{ij}^2 + a_{ij}^2 e^{-D}} \tag{2.60b}$$

$$a_{ij}^2 = a_i a_j \tag{2.60c}$$

$$D = \frac{\mathbf{r}_{ij}^2}{4a_{ij}^2} \tag{2.60d}$$

where the Coulombic interaction between to atomic partial charges, is combined with the Born equation, by a function f_{ij} which depends on the atomic charge distance, $\mathbf{r}_{ij}$, and the Born radii for both atoms, a_i and a_j. Similarly to non-bonded neighbour lists, the a_i and a_j values do not have to be updated so often, as the the dependence on other atoms is relatively weak.

The Poisson-Boltzmann equation can be replaced with the generalised Born equation to give another method termed, Molecular Mechanics and the Generalised Born Surface Area approximation (MMGBSA).

Solvent accessible surface area

The non-polar contribution to the free energy of solvation, $\Delta G^{sol}_{non-pol}$, is comprised of the the van der Waals interactions involved in the solvation process, $\Delta G^{vdW}_{non-pol}$, and the free energy required to form a cavity for the protein-ligand complex, in the aqueous solution, $\Delta G^{cav}_{non-pol}$. Forming a cavity in the solvent is an energetically unfavourable process and so $\Delta G^{cav}_{non-pol}$ will be positive. The $\Delta G^{vdW}_{non-pol}$ is favourable, and will thus have a negative value. The $\Delta G^{sol}_{non-pol}$ is determined like so:

$$\Delta G^{sol}_{non-pol} = \Delta G^{cav}_{non-pol} + \Delta G^{vdW}_{non-pol} = \gamma A + \beta \tag{2.61}$$

where A is the total surface accessible surface area (SASA), and γ and β are constants derived empirically. The linear relationship between SASA and $\Delta G^{sol}_{non-pol}$ is explained using two assumptions. First, it is assumed the solvent molecules most effected by the formation of a cavity and redistribution around the molecular surface is the first solvation shell, and thus the non-polar interactions are proportional to the SASA. Secondly, due to the rapid decay of the van der Waals potential energy, the solute-solvent van der Waals interaction are, again, assumed to be proportional to the SASA.

Calculating configurational entropy

Although some consideration has been given to entropic penalties, namely the entropic penalty of creating a cavity in solvent, for the solute, there has been no discussion about the configurational entropy. This describes the changes in the configurational degrees of freedom, of the ligand and protein, once the two solutes are associated. There is an entropic cost to ligand-protein association, because once bound, both entities are restricted in the number of conformations available to them. This can be incorporated into the free energy estimate, obtained via the

MMPBSA method:

$$\Delta G_{bind} = \Delta G_{bind}^{MMPBSA} - T\Delta S_{conf} \tag{2.62}$$

where ΔG_{bind}^{MMPBSA} is the binding free energy estimate achieved using MMPBSA, and $T\Delta S_{conf}$ is the configurational entropy as a function of temperature, T. To determine ΔS_{conf} then, the entropy S, needs to be determined for each species, and is subsequently calculated as seen in Eqn. 2.63a.

$$S_{conf}^{X} = S_{conf}^{vib} + S_{conf}^{rot} + S_{conf}^{trans} \tag{2.63a}$$

$$\Delta S_{conf} = \Delta S_{conf}^{com} - \Delta S_{conf}^{rec} - \Delta S_{conf}^{lig} \tag{2.63b}$$

The configurational entropy is the sum of the vibrational, rotational and translational degrees of freedom (Eqn. 2.63b), where S_{conf}^{X} is the configurational entropy for any of the three components. The entropic components on the right side can be determined using statistical mechanics expressions.

A common approach in calculating configuration entropies is via normal mode analysis (NMA, [84]), which achieves relatively converged entropy values. The idea is that the terms on the right side of Eqn. 2.63b are estimated using the frequencies from NMA. Conventionally, NMA is performed by removing all explicit solvent molecules, and truncating the protein to a region around the ligand binding site. This is because of the high computational demands, and difficulty in obtaining converged entropies, when a fully explicit system is used. This truncated configuration is minimised and harmonic frequencies are calculated.

There are a number of limitations with using NMA to predict entropies. First, is the computational demands associated with performing these estimations. Second,

is that estimates are calculated from minimised structures in the absence of solvent, which is not representative of the hydrated simulation. Finally, NMA estimates frequencies from a single minimum, which again is not representative of an energy landscape which contains many minima. Other quasi-harmonic approaches [85] have been applied, with unconvincing results.

A SASA-based method for the estimation of configurational entropy has been developed [86]. This approach has the benefits that it estimates entropy contributions based of actual snapshots, inclusive of water, and does not require minimisation or truncation of the system. Further, it can be completed on conventional desktops. A theoretical description of this approach can be found in section 5.1.2.

Multi-trajectory methods and adaptation energy

Binding free energies, using the MMPBSA method, can be calculated using either single or multiple simulations. Usually, free energies for complex, protein and ligand are extracted from a single simulation of the complex, termed 1-trajectory (Eqn. 2.64a). This approach has the advantage that the energies which are not involved in ligand binding, namely the internal bonded interactions, ΔG_{int}^{MM}, cancel out exactly, and so have lower computational requirements.

Alternatively, one may calculated ΔG_{bind} from independent simulations of the complex and receptor, where the free energy of the ligand is obtained from the complex simulation. This is termed the 2-trajectory approach (Eqn. 2.64b). Lastly, free energies can also be obtained from three separate simulations of each component, and is called the 3-trajectory method (Eqn. 2.64c).

$$\Delta G_{bind}^{1-traj} = G_{com}^{com} - G_{rec}^{com} - G_{lig}^{com} \tag{2.64a}$$

$$\Delta G_{bind}^{2-traj} = G_{com}^{com} - G_{rec}^{rec} - G_{lig}^{com} \tag{2.64b}$$

$$\Delta G_{bind}^{3-traj} = G_{com}^{com} - G_{rec}^{rec} - G_{lig}^{lig} \tag{2.64c}$$

In the 1-trajectory approach the free energy estimate of the ligand will be limited to the conformations that are sampled by the protein within the complex trajectory. But in practice, a ligand can cause conformational changes in the protein, and vice versa, known as the adaptation energy.

Adaptation energies allow one to understand the energetics involved in ligand binding. For example, a small adaptation energies for both the receptor and ligands would suggest a 'lock-and-key' method of binding, where large changes in adaptation energy for receptor, and negligible difference for the ligand, would follow the 'induced-fit' mechanism.

If one averages the G_{rec} value obtained from the 1-trajectory method, for all ligands, then the difference between the original G_{rec} value and the averaged value is an estimated adaptation energy of the receptor. Similarly, evaluating G_{lig} from free ligand MD trajectories, allows one to compare the ligand adaptation energy. This is the difference between the G_{lig} obtained from conformations bound to the protein, and the G_{lig} from the free ligand simulation.

2.7 Ensemble-based binding affinity predictions

The basis of all studies in this thesis, is the use of ensemble-based binding affinity predictions, which was briefly explained in section 2.3.4. In the following sections, a more detailed explanation will be presented of two ensemble-based approaches, namely Enhanced Sampling of Molecular Dynamics with Approximation of Continuum Solvent (ESMACS, [16]), and Thermodynamic Integration with Enhanced Sampling (TIES, [42]).

2.7.1 Enhanced Sampling of Molecular Dynamics with Approximation of Continuum Solvent

ESMACS computes 25 identical simulations, of a specific ligand and receptor complex, with the only difference being the initial velocities assigned to the atoms after minimisation, according to a Maxwell-Boltzmann distribution at constant temperature. This group of 25 simulations is termed an ensemble and the individual simulations themselves are replicas.

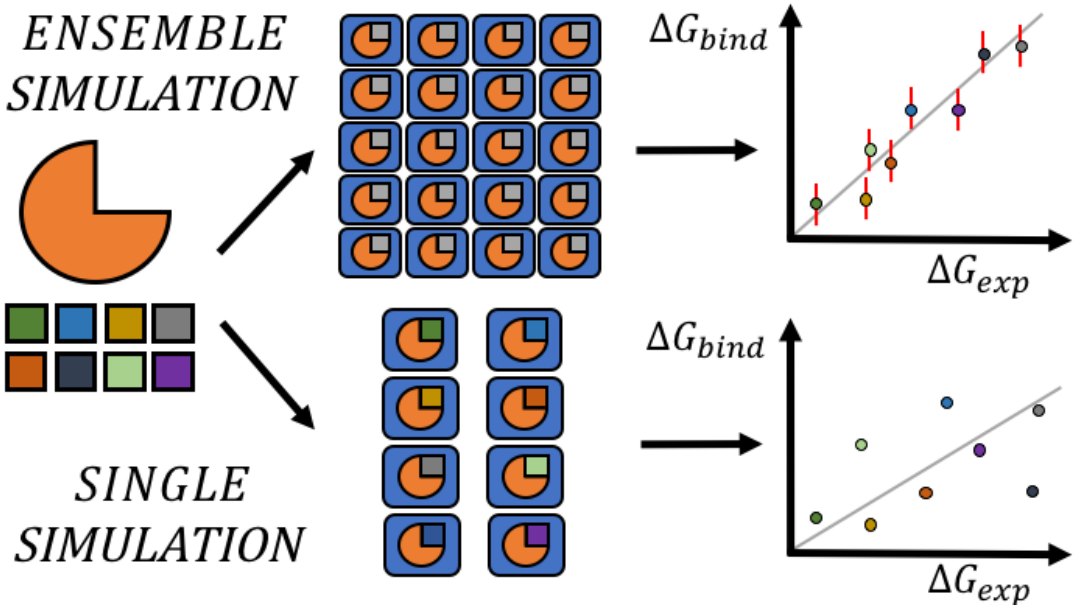
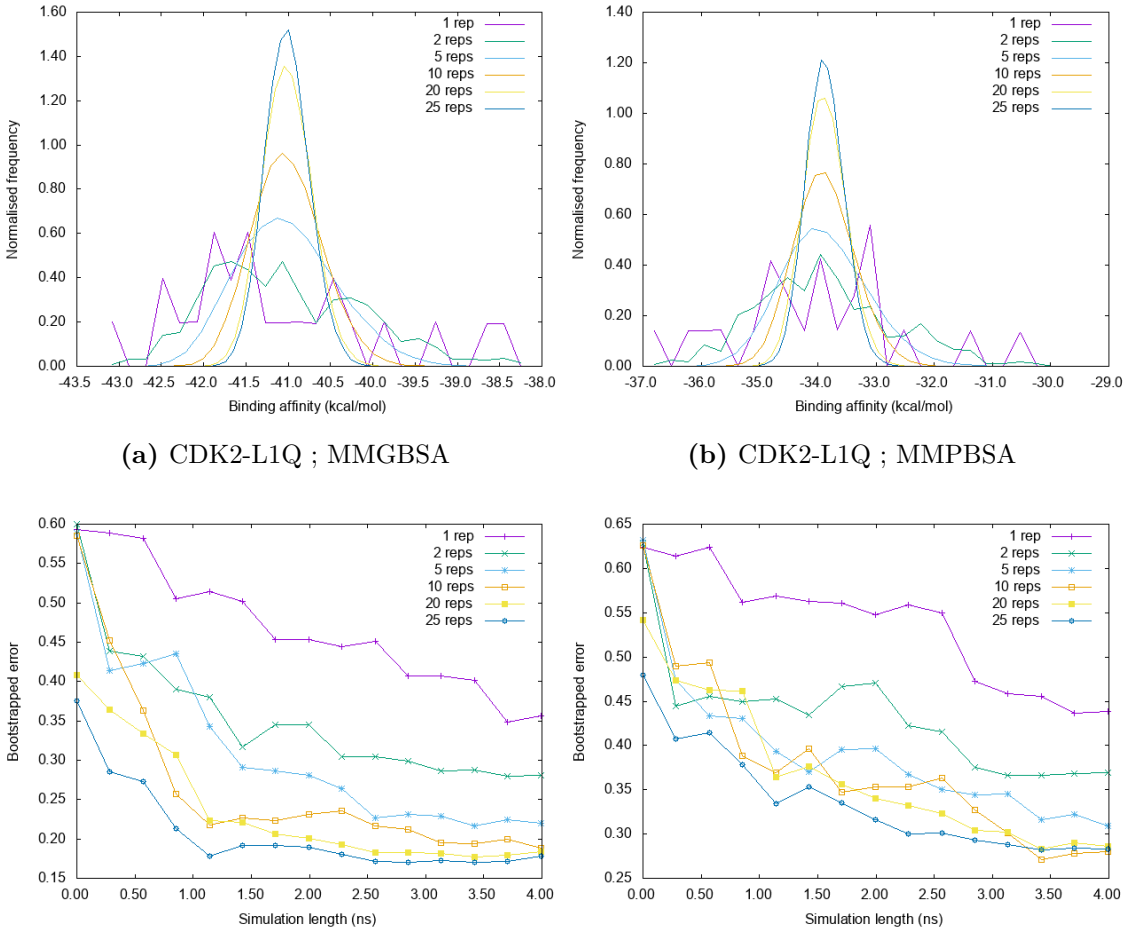


Figure 2.6: *A schematic representation of ensemble simulation, performed in ESMACS, and single simulations. Running multiple replicas allows for tight control of errors, and so we obtain reliable and reproducible binding affinity predictions. Single simulations do not have error bars, and thus yield unreproducible results.*

Upon completion of the ensemble simulations, MD trajectories are used to compute the free energies are calculated based on the extended MMPBSA and MMGBSA method described in section 2.6.2.3, for each replica. We have witnessed that the frequency distribution of these set of values lie essentially on a Gaussian curve (Fig. 2.3). A number of statistical techniques are applied to produce a mean final ΔG value, and the standard deviation associated with it. This will be described in

the next section.

Structuring the MD simulations and free energy calculations in such a way, allows for reproducible results, and thus makes it reliable. The strength of this approach is that it allows for precise ranking of drugs to a protein target.



(a) CDK2-L1Q ; MMGBSA

(b) CDK2-L1Q ; MMPBSA

(c) TYK2-L02 ; MMGBSA

(d) TYK2-L02 ; MMPBSA

Figure 2.7: *Plot of the variation of the bootstrapped statistics as a function of replica number and simulation length: (a) and (b) show the variation and normalised frequency of the error as a function of replica number for the CDK2-L1Q complex using the MMGBSA and MMPBSA method, respectively; (c) and (d) show the varation of the bootstrapped error as a function of simulation length for the TYK2-L02 complex, using the MMGBSA and MMPBSA, respectively. The above benchmarking tests justify the replica number and simulation length selection in ESMACS.*

Two factors determining the sampling of phase space, in ESMACS, are the number of replicas in an ensemble simulation, and the simulation length. Both need to be gauged to ensure that a sufficient amount of phase space is sampled, but equally the wall clock time remains reasonably low. Benchmarking tests have shown that 25 replicas and an equilibration length of 2 ns, followed by 4 ns of MD simulation [43], achieves statistically converged binding affinities. Fig. 2.7 presents the quantifies the variation of error as a function of replica number and simulation length for two receptor-ligand complexes that are investigated in chapter 4.

2.7.2 Thermodynamic Integration with Enhanced Sampling

In TIES, 5 replicas of 13 lambda states are generated, for which $\Delta V/\Delta\lambda$ are calculated, that is 65 simulations in total. This was justified through benchmarking tests, which showed that this number of replicas, and λ windows gave the best compromise between accuracy and reproducibility in $\Delta\Delta G$ values, and computational expense [42]. Potential derivatives for each replica are the average potential energy derivative over the entire trajectory. $\Delta V/\Delta\lambda$, for each lambda, is calculated by averaging the 5 potential derivative values obtained. Then the integral in Eqn. 2.48 is solved numerically, using the trapezoidal rule, to achieve ΔG^{aq}_{alch} and ΔG^{bound}_{alch} for the free ligand and complex, respectively.

It should be noted that TIES is not limited to these parameters. If one wishes to increase phase space sampling, then system-specific modifications to the replicas size, simulation length or number of λ values can be made. Error quantification analysis was performed by Bhati et al. [42] to justify the selection of 5 replicas per λ window and simulation length of 4 ns, reporting similar results to what is seen in Fig. 2.7.

Binding free energies obtained by TIES is a stochastic variable described by a Gaussian random process, which means that the integral in Eqn. 2.48, is itself a

stochastic variable (described in 2.3.4). Therefore, statistical mechanics permits us to interpret Eqn. 2.48 in terms of stochastic calculus. In short, the integral that is computed for each λ window is for the average of 5 replicas, which consist of MD trajectories that also display random Gaussian distributions. This allows for a more reproducible calculation of each λ integral and effectively 'smoothes' the integral that is computed. The conventional approach would be to run a single replicas for all λ windows, and subsequently compute the integral for all λ. This is then repeated n times and then an average is taken. The difference here is that single potential derivative value can lie anywhere within, what would be, large error bars. As a result, n set of integrands with a 'jagged curve' from which an inaccurate, and unrepeatable ΔG_{alch} is generated.

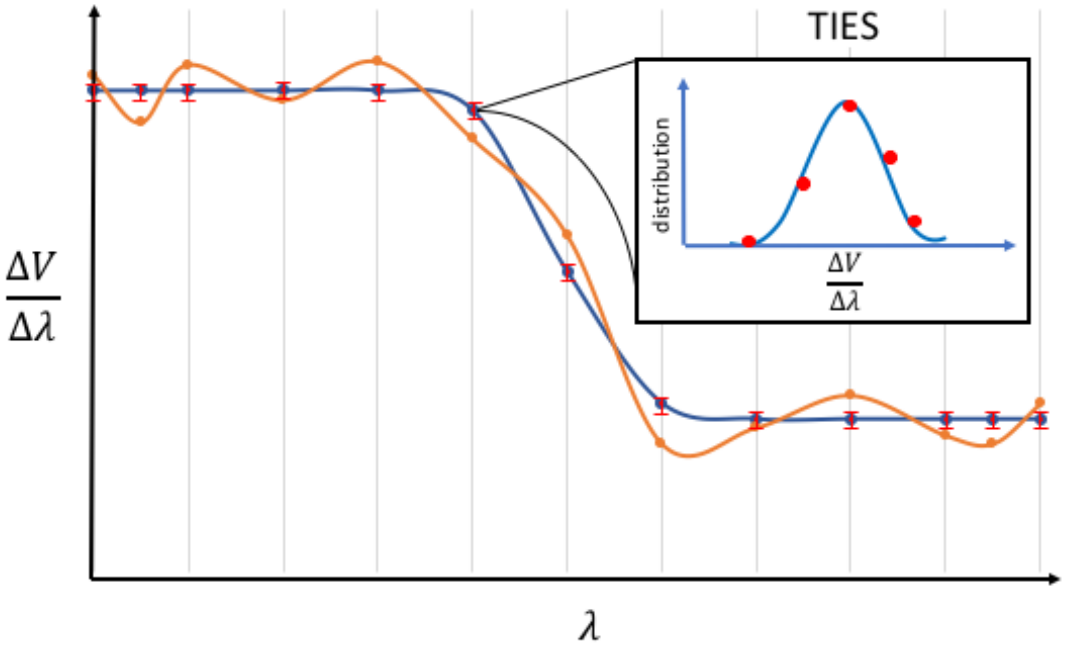


Figure 2.8: *A representation of ensemble-based free energy methods, compared with single simulations. TIES averages $\Delta V/\Delta\lambda$ from all replicas, with respect to λ; which gives a close control over. This is depicted by reproducible blue line, with small red error bars. The integral is then numerically calculated from the resultant averages. Conversely, evaluations of $\Delta V/\Delta\lambda$ from single simulations lead to largely varied results which change with each new simulation. This is represented with an orange line, and an absence of error bars.*

In the same way, we are able to use to stochastic calculus to generate standard errors. The error for each λ window is the bootstrapped standard error of the mean for all potential derivative, of all replicas. The variance of ΔG^{aq}_{alch}, ΔG^{bound}_{alch} and the final relative binding affinity are as follows:

$$\sigma^2_{aq/bound} = \sum \sigma^2_{\lambda} (\Delta\lambda)^2 \tag{2.65a}$$

$$\sigma^2 = \sigma^2_{bound} + \sigma^2_{aq} \tag{2.65b}$$

Hysterisis, the sum of $\Delta\Delta G_{bind}$ associated with the closing of a thermodynamic cycle, in theory, is zero. Due to the finite integration of the potential derivative, uncertainties are inevitable, and so hysteresis in practice is never zero. Thus, as the calculated hysteresis tends to zero the accuracy of the predictions improves. Theoretically, the difference of the integral of the forward and reserve transformation is also zero. Through the use of stochastic calculus (section 2.3.4) in the aforesaid fashion, in conjunction with replica simulations per lambda window, TIES keep hysteresis to a minimum. This is achieved through sufficient sampling of phase space and the correct approaches taken towards error propogation. Recent methods [46, 45] have reported low hysterisis by averaging total hysteresis over a full cycle closure, generating unrepresentative errors. The same method reports relative binding free energies from single replica simulations, and with no standard error assigned to the calculated $\Delta\Delta G$.

Chapter 3

Methods

In this chapter, a detailed account of the ESMACS and TIES methodology is presented. Then, it is described how the methods are automated and coupled with high-performance and cloud computing, through the application of the Binding Affinity Calculator (BAC, [87]).

3.1 Enhanced Sampling of Molecular Dynamics with Approximation of Continuum Solvent

We begin with ESMACS, the approach based on ensemble MD simulations that estimates binding affinities using end-point free energy methods. The theoretical background behind this protocol is found in section 2.7.1.

3.1.1 Model preparation

Geometry optimisation of the ligands was completed using Gaussian03 [88], and the restrained electrostatic potential (RESP, [89]) method was used to generate partial atomic charges. Ligand parameters were created using the general AMBER force field (GAFF, [67]). GAFF is used as it can generate a force field model automatically and is compatible with the AMBER protein force field. With regards

to the protein parameters, ff14SB [90] were loaded, along with the parameters for phosphorolated threonine (T2P, [91]) and phosphorylated serine residus (S2P, [91]).

The Leap module in AMBER14 [89] was used to electrically neutralise the complex, using counter ions, and solvated in a cubic box with a 14 Å buffer, using atomistic TIP3P water [92]. Topology and coordinate files were subsequently created.

3.1.2 Simulation set-up

Equilibration and subsequent MD production runs were performed by the MD package NAMD2.9 [93] using an isobaric-isothermal (NPT) ensemble. NAMD2.9 uses the velocity Verlet integration method to advance positions and velocities through time. The SHAKE algorithm [60] was included for all atoms covalently bonded to hydrogen, allowing for an integration time step of 2 fs. Periodic boundary conditions were used through equilibration and simulation stages. The particle mesh Ewald summation method (PME, [58]) was used to handle long-range Coulombic interactions, with a cut-off distance of 12 Å. Each equilibration and production stage was performed in replicates of 25.

Minimisation was conducted using the conjugate gradient and line search methods for 2000 iterations whilst achieving a gradient tolerance of 10. The system was then annealed from 50 K to 300 K, over a period of 50 ps, where it was maintained at 300 K, and 1 bar using the Langevin thermostat and the Berendsen barostat, respectively. Equilibration phase was 200 ps while maintaining a 4 kcal/mol $Å^2$ force constant on both ligand and receptor, to ensure solvation of the complex. This was followed by step-wise force constant relaxation for both the ligand and receptor. Firstly, the ligand force constants were reduced, in increments of 1 kcal/mol $Å^2$, from 4 to 0 kcal/mol $Å^2$ over a period of 200 ps. The receptor force constants were similarly reduced from 4 to 1 kcal/mol $Å^2$ for 150 ps. Finally, all constraints were removed and the system was equilibrated, unrestrained, for 200 ps.

MD simulation were performed over a period of 4 ns and coordinates were recorded every 1 ps. The simulation length was decided based on previous studies which showed sufficiently converged free energy values at this time length [16, 43].

3.1.3 Free energy calculation

The binding affnities were achieved using the MMPBSA and MMGBSA method via the AMBERTools15 [90] package. MMPBSA.py.MPI [94] module was used for the calculations. The module employs SANDER to calculate the bonded and non-bonded molecular mechanics terms E_{bnd}, E_{elec} and E_{vdw} with no cut-off assigned for non-bonded energies.

The G_{pol} is described by the PB or GB equation (see section 2.6.2.3). Both methods assign an internal and external dielectric constant of 1 and 80, respectively. For PB [72], the linear PB equation was solved on a cubic lattice with 0.5 Å grid spacing. The GB model is described by Onufriev et al. [83]. In the case of MMPBSA G_{nonpol} term, $\gamma = 0.0052$ kcal/mol Å and $\beta = 0.92$ kcal/mol, and for MMGBSA, γ and β were set to 0.0072 kcal/mol Å and 0.92 kcal/mol, respectively.

Both ΔG (via PB and GB methods) and ΔTS were calculated using 48 snapshots, out of a possible 400, from the 4 ns simulation trajectory. The snapshots were extracted evenly i.e. every 8th snapshot. ESMACS calculates the binding free energy obtained from the PB/GB method alone, and with the inclusion of configurational entropy and free energy of association. For each method, free energies are averaged over 48 snapshots to get a free energy per replica. To obtain the final binding free energy for a ligand-protein system, an average is taken of all 25 replicas.

3.1.4 Statistical analysis

To calculate mean ΔG_{bind} we use all 1200 binding free energies generated from the 48 frames of 25 replicas. The mean ΔG_{bind} is evaluated by statistical bootstrapping

techniques, the data is re-sampled with replacement, 100,000 times. This means that there is a possibility that a data point can represented more than once, and others may never be sampled. Standard errors are the standard deviations of the mean ΔG_{bind}. The 3-trajectory method generated free energy values of the complex, receptor and ligand from individual MD simulations, and so the bootstrapping protocol described above performed for each trajectory.

3.2 Thermodynamic Intergration with Enhanced Sampling

The TIES protocol also adopts ensemble simulations and calculates the relative binding affinity, or free energy difference, between two ligands. A theoretical description of this method is presented in section 2.7.2.

3.2.1 Model preparation

TIES follows an identical process as ESMACS for geometry optimisation and parameterisation of the individual ligands. However, the dual topology method [95] employed by TIES means that hybrid PDB structures, parameters and topologies need to be created specifically for each alchemical ligand transformation.

Firstly, the coordinates from the ligand PDB structure are overlapped for ligands at $\lambda = 0$ and $\lambda = 1$ (that is the initial and final ligand) for the common region found in both states. To qualify as the 'common region', the atom charge must not be greater than 0.1 e between the equivalent atoms on each ligand. This criterion, however, can be modified for highly charged, or polar ligands. The partial charges for each atom in the common region is the average partial charge of the equivalent atom in each ligand. Following this, the partial atomic charges and parameters were generated for the hybrid ligand, using the same approach seen in ESMACS. Protein parameters were achieved in the same way as ESMACS.

With regard to model preparation, the only difference here between TIES and ESMACS is that TIES contains the hybrid ligand PDB structure bound to thes receptor and not the individual ligands. From this point, Tleap, as in ESMACS, is employed to build the topologies and solvate the complex.

3.2.2 Simulation set-up

Equilibration and simulation stages are identical in the duration, ensemble selection (NPT), and conditions at which simulations were conducted, as ESMACS. Although almost all simulation parameters are shared between the two methods, TIES requires some additional selections because of the dual topology method that has been employed.

A bounded (soft-core) van der Waals (vdW) potential [96, 97] was used (default coefficient of 5) to ensure the vdW potential is finite across the whole perturbed space, and avoids overlapping of atoms at low λ values (so-called "end-point catastrophes"). Alchemical decoupling was turned on so that only non-bonded interactions were scaled and the interactions within the perturbed region was preserved. Electrostatic and vdW interactions were scaled separately. For the created atoms, the electrostatic interactions were fully coupled at $\lambda = 1$ and fully decoupled between $\lambda = 0$ and $\lambda = 0.45$. Between $\lambda = 0.45$ and $\lambda = 1$, the electrostatic interactions were linearly coupled. Conversely, for the annihilated atoms, the electrostatic interactions were fully coupled at $\lambda = 0$ and fully decoupled between $\lambda = 0.55$ and $\lambda = 1$. Between $\lambda = 0$ and $\lambda = 0.55$ the annihilated atoms were decoupled linearly. Van der Waals interactions were linearly decoupled from $\lambda = 0$ to $\lambda = 1$ for created atoms, and vice versa for annihilated atoms.

13 λ windows were implemented in TIES. This is composed as follows: $\lambda = 0.0$, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 1.0. Simulations were performed, in replicates of 5 per λ window. Coordinates were recorded every 10 ps, and $\Delta V/\Delta\lambda$ every 2 ps.

3.2.3 Statistical analysis

The error analysis in TIES is performed similarly to ESMACS. The bootstrapping technique described in section 3.1.4 is performed for each λ window, that is 10000 $\Delta V/\Delta\lambda$ recordings corresponding to 2000 recordings for every replica. This data set is re-sampled with replacement, 10,000 times for each λ window from which mean and standard deviations are calculated.

3.3 Binding Affinity Calculator

A distinguishing feature of the TIES and ESMACS methodology compared to other binding affinity methods [45, 46, 44, 98], is the high level of automation. TIES and ESMACS are facilitated by the Binding Affinity Calculator (BAC, [87]) through an automated workflow. BAC is an e-infrastructure tool that comprises a selection of software programmes and services. It automatically completes the model building stage of both protocols, this is followed by submitting large numbers of replica simulations to high performance computers (HPCs) or in the cloud, and subsequently collecting and analysing the data to evaluate free binding energies, and errors that are associated with this. BAC eradicates a vast amount of manual overhead, which naturally leads to removal of human error.

FabSim [99], a toolkit integrated within BAC, completes a range of computational tasks. It is responsible for creating a uniform directory structure, which makes it easy for users to navigate through input and output files, and transport the data across different machines. FabSim also submits replica jobs to HPCs or clouds of choice. Within this, input files are automatically copied to the remote machine and, upon completion of calculations, the subsequent files are transferred back to the user's local machine.

Completing complex workflows, like TIES and ESMACS, requires a user to have a high technical ability. The development of a user-friendly BAC (uf-BAC, [42, 100])

will allow for non-technical users to calculate binding affinities. This is targeted at medicinal chemists and biochemists who would like to support their experimental results with calculated values.

3.4 High-performance and cloud computing

Ensemble-based free energy predictions, like ESMACS and TIES, are not possible without utilising industrial strength computers, namely HPC and cloud services. Table 2.1 demonstrates the vast computational resources required to perform ESMACS and TIES calculations for one ligand-protein complex, or a single alchemical transformation, respectively.

Consequently, the two ensemble-based approaches have been used on an unprecedented scale, when over 50 protein-ligand complex and perturbations were studied, resulting in considerable news coverage [101, 102]. To achieve this level of scientific output, approximately 250,000 cores were required, for an uninterrupted duration of 36 hours. Thus, very significant resources, and wall clock time, are required to generate binding affinity predictions on industrially relevant timescales.

HPC services have their drawbacks. The process to acquire computing time involves time-consuming proposals, where compute time is allocated in yearly cycles, or longer time frames. This means users are constantly tied to the allocated computational resources, and must carefully plan projects well in advance. In the pharmaceutical environment, this framework would not be conducive to fast moving drug discovery programmes that require, on occasions, spontaneous use of compute time. Outsourcing to HPC services is often associated with queuing times for calculations to begin, which can sometimes last for several days. This too would not be acceptable in an industrial setting.

As a consequence performing 'on demand' calculations has become an attractive alternative, and cloud computing offers this service. Cloud computing is an al-

Table 3.1: *HPC requirements for ESMACS and TIES. The core counts and subsequent wall clock times are obtained from runs on the LRZ SuperMUC Phase 1 and Phase 2 machines. ESMACS/TIES calculations can be run in parallel making largely scalable dependent on the user's needs. An increase in core count is directly proportional with a speed-up in wall time. The bottle neck is normal mode analysis which is largely variable in time. Total core hour allocation for for ESMACS is calculated using an average 17 hour normal mode analysis calculation.*

Method		ESMACS	TIES
Replicas per complex		25	65
Equilibration	Location	HPC	HPC
	Cores	4,608	12,480
	Time (hrs)	1.41	2.15
Production	Location	HPC	HPC
	Cores	4,608	12,480
	Time (hrs)	2.81	4.27
Free Energy Calc. (Normal Mode)	Location	HPC	-
	Cores	1,200	-
	Time (hrs)	0.1 (9-24)	-
Statistical Analysis	Location	desktop	desktop
	Time (hrs)	0.5	0.1
Total Core Hours (approx)		40,000	80,000

ternative framework which allows users to run applications from remote resources. Access to computer time is provided in return for monetary payment. There are two modules used in cloud computing. First, there is the 'Infrastructure as a Service' (IaaS) model, which provides access to computing time, memory and storage, but the user is required to run their own applications. The other model is 'Software as Service' (SaaS), which provides software that users are able to exploit. Currently, BAC is being implemented into the SaaS model, through DNANexus, Microsoft

Azure and Amazon Web Services (AWS) cloud platforms.

Historically, central processing units (CPUs) have been used as the workhorse to perform programming tasks, however relatively recent developments in the graphics processing unit (GPU), has seen it gain traction. The use of GPUs has shown increased performance compared to CPUs, when completing MD molecular simulations, however, most GPU codes are yet to scale beyond a single node. A recent paper reports fast, accurate GPU-accelerated binding affinities, via TI calculations [98]. The method also acknowledges the requirement for ensemble simulations, and implements this approach to generate reproducible binding affinities.

3.4.1 Specification of HPC resources

The binding affinity values reported in this thesis are made possible through access to several HPCs in the UK, and Europe. Below is a description of the HPC resources that have been used, and some detail about each.

The Hartree Centre, an institution run by the the Science & Technology Facilities Council (STFC), boasts an iDataPlex and Xeon Phi, and NextScale Cluster, comprising 2,016 and 8,640 cores, respectively [103]. Both machines have infiniband interconnect. Of the 84 nodes in the iDataPlex machine, 42 nodes have accelerators. The de-commissioned BlueGene/Q (98,304 cores) and BlueGene/Q BGAS (40,960 cores) machines were also used.

ARCHER, the UK National Supercomputing Service, is based around a Cray XC30 supercomputer which contains 4,920 nodes (9,840 cores) [104]. ARCHER has compute nodes that have 64 GB memory shared between two cores, and a small number of high-memory nodes (128 GB). ARCHER uses the Cray Aries interconnect to link all nodes.

SuperMUC Petascale System is the name of the HPC at the Leibniz-Rechenzentrum (Liebniz Supercomputing Centre, LRZ) near Munich, Germany [105]. SuperMUC

has more than 241,000 cores at its disposal. The phase 1 installation is made up of three clusters: BladeCenter HX5 (8,200 cores), and two iDataPlex dx360M4 machines (147,456 and 3,840 cores, respectively). The phase 2 installation is a NeXtScale nx360M5 WCT with 86,016 cores.

Chapter 4

Application of ESMACS and TIES in the context of a drug discovery programme

4.1 Introduction

Recently, there have been promising advancements in MD-based free energy protocols aimed at achieving results on an industrially relevant timescale. Wang and colleagues [46, 45] – researchers at Schrödinger – reported good relative binding affinity predictions, using their FEP+ methodology. A strong correlation was seen with experimental results, for 8 different biological systems, and anti-correlated binding affinities were obtained using the MMGBSA free energy method. Similarly, Aldeghi and colleagues [44] performed FEP-based absolute binding free energy calculations on a diverse range of ligands. Good correlation with experimental binding affinities, and tight error bars were reported, however, this approach is so far limited to rigid drug targets only.

Coveney and colleagues have developed two ensemble-based approaches, utilising both alchemical and end-point methods in each case. Running ensemble simulations

in parallel on high-performance computers (HPCs), and in the cloud, give rise to rapid, accurate and precise binding affinity predictions. These are termed "Thermodynamic Integration with Enhanced Sampling" (TIES, [42]) and "Enhanced Sampling of Molecular dynamics with Approximation of Continuum Solvent" (ESMACS, [16]). The frameworks, and the associated theory that underpins these approaches, are described in detail in chapter 2 and 3.

These two approaches have been used to estimate binding affinities for a selection of ligands, across 5 different receptor targets, which have been adopted from Wang et al. [46, 45]. In this chapter, we will compare and contrast the TIES and ESMACS approach with that of Wang et al [45]. TIES results for this work have recently been published [42] and so the first half of this chapter will report results obtained using the ESMACS method. We have obtained binding affinities for 100 ligands across five drug targets.

The latter part, will address the TIES results where 61 ligand transformations where analysed. Although TIES is able to yield more accurate results, the computational resources required are vast compared to ESMACS. It would be beneficial, then, to be able to run TIES calculations using a smaller number of λ windows. Here we investigate if this is at all possible. The overarching motivation of this chapter is to gain knowledge about how TIES and ESMACS perform across these 5 systems, and if one approach could be used over the other, in each case.

4.1.1 Biological activity and role in disease

The systems that are studied in this chapter have been selected for three reasons. Firstly, this chapter compares TIES and ESMACS with another methodology [46, 45] that determines ligand binding affinities. For this reason, we maintained the same systems that were used in that study, achieving a direct comparison in this respect. The receptor and ligand structures have been extracted from previous publications which are outlined in Fig. 4.1 and Table 4.1.

Secondly, the systems studied cover a number of different protein families (i.e. kinase, phosphotase, protease) which vary in size, mechanism of binding and biological function. The ligands for which binding affinities are calculated also vary in size, charge and flexibility. Investigating ESMACS and TIES across a diverse range of proteins and ligands will give a good indication of the robustness of these methodologies. Ultimately, such methods are being developed for the application in the drug discovery setting, and therefore need to produce reliable binding affinities regardless of the system of study. Lastly, these proteins play an important role in a number of disease areas which are current areas of interest. The biological activity and role in disease are described below.

Cyclin-dependent kinase 2

Cyclin-dependent kinases (CDKs) are a member of a family of enzymes that regulate cell proliferation in the eukaryotic cell cycle. The cell cycle is the process in which a cell is divided and duplicated to produce two replica daughter cells. CDK2 regulates the initiation of the DNA synthesis during interphase [106].

CDK2 activity has been described by Russo et al. [107]. CDK2 is fully activated by a two-stage process: firstly, it binds to regulatory sub-units called cyclin A, which induces low level catalytic activity [107]. CDK2, now complexed to cyclin A, is phosphorylated by a CDK2-activating kinase (CAK) at a conserved threonine residue located on the CDK2 regulatory T-loop. CAK is constitutively active and so phosphorylated CDK2 concentrations are regulated by the presence of cyclin A. The loss of CDK2 activity can result in loss of G1 check-point control and subsequently gives rise to unregulated cell growth [108]. CDK2 has been a popular drug target because they are directly involved in the regulation of cell proliferation.

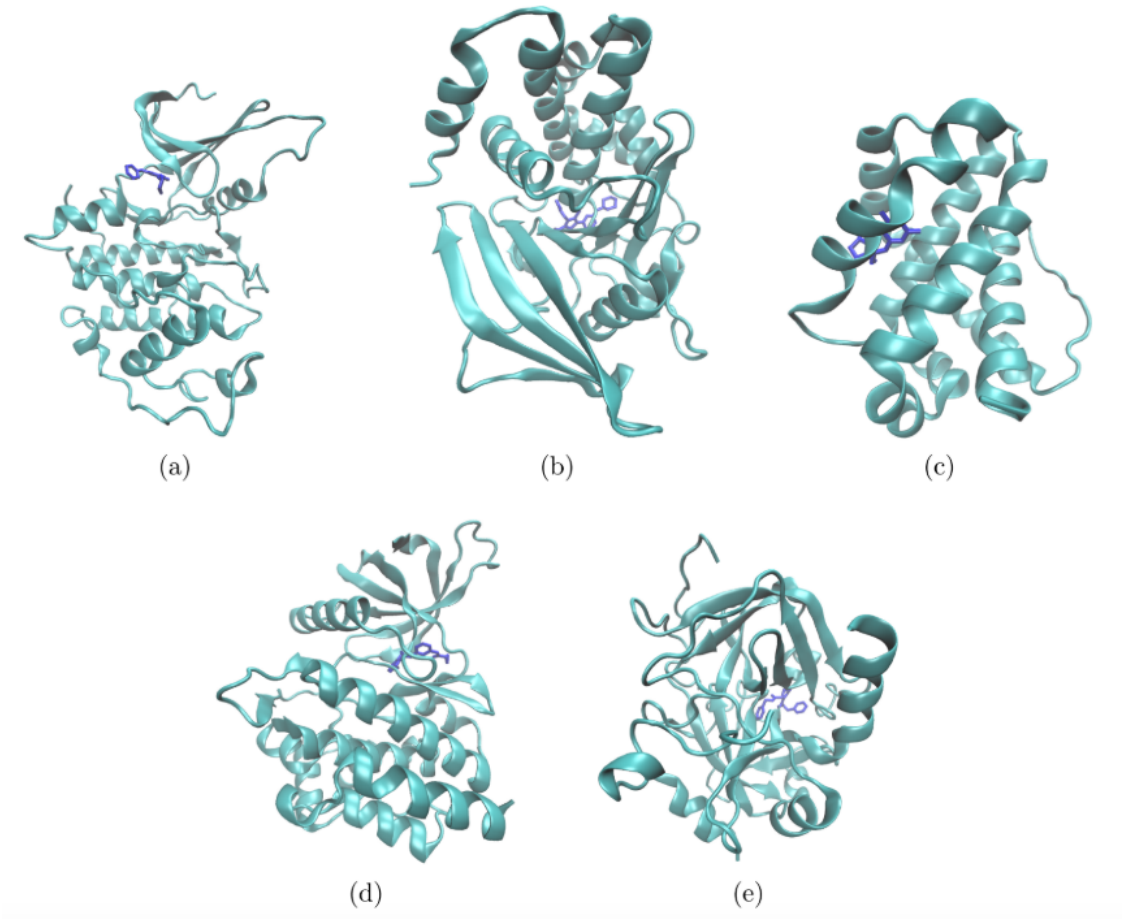


Figure 4.1: *Structures of the 5 receptors used in this chapter shown as a teal ribbon representation: (a) cyclin-dependent kinase 2 (CDK2); PDB code: 1H1Q, (b) protein tyrosine phosphotase 1B (PTP1B); PDB code: 2QBS, (c) induced myeloid leukemia cell differentiation protein 1 (MCL1); PDB code: 4HW3, (d) non-receptor tyrosine-protein kinase 2 (TYK2); PDB code: 4GIH, (e) thrombin; PDB code: 2ZFF. Each receptor is shown with a ligand (blue stick representation) present in the binding pocket. Additional physical and structural properties are presented in Table 4.1.*

Non-receptor tyrosine-protein kinase 2

Non-receptor tyrosine-protein kinase 2 (TYK2) is one of four Janus kinases (JAKs) – the others being JAK1, JAK2 and JAK3 – and is associated with cytokines and growth factor proteins in mediating inflammation [109]. JAKs interact with a specific set of receptors, which ultimately results in the creation of docking sites on

signal transducers and activators of transcription (STAT) proteins. STAT proteins are phosphorylated by JAKs and mediate gene transcription.

TYK2 interacts with the IL-12/IL-23 pathway [110, 111] which are associated with the T helper type 1 (Th1, [112]) and T helper type 17 immune responses (Th17, [113]). TYK2 phosphorylates the STAT proteins which then transolacates into the nucleus and mediates gene expression. The pathogenesis of psoriasis and inflammatory bowel disease (IBD) is linked to abherrent function of the Th1/Th17 immune responses and IL-12/IL-23 pathways, resulting in inflammation of the skin and gut, respectively [114, 115].

Induced myeloid leukemia cell differentiation protein 1

Normal, non-cancerous cells that exhibit aberrant growth are subject to programmed cell death [116, 117]. A dysfunctional induced myeloid leukemia cell differentiation protein (MCL1), has been found to circumvent programmed cell death of cancer cells [118]. It has also been found that aberrant function of MCL1 is one of the most common traits found in human cancer [119, 120]. Studies that silenced the MCL1 gene show to significantly decline particular non-small-cell lung cancer (NSCLC, [121]), suggesting that MCL1 is has potential to be an effective drug target.

Thrombin

Thrombin is a serine protease that cataylses the formation of fibrin by cleaving the peptide bond in fibrinogen. This increased concentration of fibrin activates fibrin stabilising factor 13 (Factor XIII) and results in platelet aggregation and formation of thrombus [122]. Although this function is extremely important – without this our blood would not be able to clot – it is the cause of dangerous intravascular clot formation, which results in several cardiovascular diseases (myocardial infarction, deep vein thrombosis and ischemic stroke, to name a few). Historically,

anticoagulants like heparin and warfarin were administered subcutaneously which for obvious reasons is undesirable, and so thrombin has long been a desired drug target which offered the benefit of oral administration.

Protein tyrosine phosphatase 1B

Protein tyrosine phosphatase 1B (PTP1B) is an attractive target for type 2 diabetes and obesity, as it is a negative regulator of the insulin and leptin pathways [123, 124]. Phosphotases have the opposite function of kinases (such as TYK2 and CDK2) and are responsible for dephosphorylation. That is, PTP1B has found to supress generation and subsequent secretion of insulin and leptin, resulting in high blood sugar levels. Studies have shown that mice without PTP1B have displayed higher sensitivity to insulin, more efficient glycemic control and resistance to obesity as a result of high fat diets [125, 126].

4.2 Methods

All models in this study were extracted from the references available in Table 4.1. Model preparation and simulation set-up were performed using the Binding Affinity Calculator (BAC) which is described in chapter 3.

All binding affinities generated in this chapter have been performed used the ESMACS and TIES methodology which is also described in chapter 3.

4.3 Results

We will critically assess the performance of ESMACS and TIES, across 5 systems of study: CDK2, TYK2, PTP1B, MCL1 and thrombin. Knowledge gained within this section will allow us to understand how these ensemble-based methods can be applied in the context of a drug discovery programme. A thorough description of TIES and its performance across the systems in this chapter, has recently been

published by Bhati et al. [42].

Table 4.1: *An overview of the systems explored in this chapter, and the predictive performance of the ESMACS protocol for all trajectory methods. GB and PB are the generalised Born, and Poisson-Boltzmann free energy methods used to obtain binding affinities.*

		CDK2		TYK2		MCL1		PTP1B		Thrombin	
No. of ligands		16		16		42		15		10	
PDB code		1H1Q		4GIH		4HW3		2QBS		2ZFF	
Structure resolution method		XRD		XRD		XRD		XRD		XRD	
Structure resolution (Å)		2.5		2.0		2.4		2.1		1.47	
Reference		[108]		[127, 128]		[129]		[130]		[131]	
ΔG_{exp} method		IC_{50}		K_i		K_i		K_i		ITC	
ΔG_{exp} range (kcal/mol)		4.21		4.28		4.18		5.13		1.7	
Receptor size (residues)		297		288		150		298		231	
Charged ligands		No		No		Yes		Yes		Yes	
		GB	PB	GB	PB	GB	PB	GB	PB	GB	PB
Pearson, r_p	1-traj	0.49	0.61	0.70	0.73	0.75	0.80	-0.41	-0.40	0.85	0.92
	2-traj	-0.55	-0.56	0.56	0.71	0.62	0.64	-0.20	0.02	0.31	0.07
	3-traj	-0.56	-0.56	0.41	0.47	0.64	0.66	-0.25	-0.04	0.23	-0.01
Spearman, r_s	1-traj	0.24	0.38	0.49	0.53	0.56	0.64	0.16	0.16	0.72	0.84
	2-traj	0.31	0.29	0.31	0.51	0.39	0.41	0.04	0.00	0.09	0.00
	3-traj	0.31	0.29	0.17	0.22	0.41	0.44	0.06	0.00	0.05	0.00
$\sigma_{average}$ (kcal/mol)[1]		0.30		0.18		0.34		0.48		0.21	

[1] $\sigma_{average}$ is the averaged standard error of all ligands within a system. $\sigma_{average}$ is reported for only the best performing ESMACS trajectory approach.

We will begin with a critique of ESMACS. Following this, a more detailed description of the systems will be presented, highlighting specific considerations that are required when completing binding affinity calculations. The subsequent segment of this chapter will compare ESMACS and TIES, working towards and understanding when each respective approach could be applied in a drug discovery setting. We close the chapter with a TIES study that explores what the lowest number of λ states is required to gain reliable results, but save on compute time.

4.3.1 An assessment of the ESMACS protocol across 5 biological systems

Table 4.1 displays a detailed description of the systems, including the predictive performance and reproduciblilty of the results. The most significant trend observed across all systems, bar PTP1B, is that the 1-trajectory method results in superior r_p and r_s values, in comparison with the 2 and 3-trajectory approaches. In fact, in the case of CDK2, we report anti-correlated results for 2- and 3-trajectory ESMACS binding affinities. The electrostatic contribution to the binding affinity is responsible for the anti-correlation seen PTP1B. This will be assessed more closely in section 4.3.3. Additionally, the PB approach is superior to GB in all systems. This is perhaps expected, as the PB approach is a more theoretically rigourous free energy method.

In all scientific work, reproducibility is of key importance in drug discovery approaches. MD trajectories diverge rapidly from differences in initial starting velocities [41, 87, 43], and as such, generating binding affinity predictions from a single trajectory gives rise to unreproducible binding affinities. ESMACS, through the use of ensembles, and the subsequent statistical bootstrapping approach, reduces uncertainty of the calculated binding affinities. Calculating standard errors has been described previously (section 3.1.4). The average of all standard error values obtained in each ligand set ($\sigma_{average}$) gives us an idea of the reproducibility of ESMACS binding affinities.

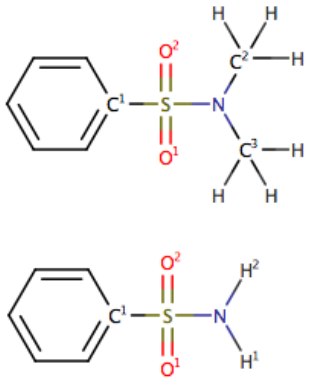
The TYK2 system, followed by CDK2, report the lowest $\sigma_{average}$ value. The standard error for any ligand in the CDK2 system is no more than 0.55 kcal/mol, and 0.21 kcal/mol in TYK2. The systems with the two lowest $\sigma_{average}$ values correspond with ligand sets that do not contain charged ligands. This therefore suggests that ESMACS loses some reliability when ligands with formal charges are assessed. The highest $\sigma_{average}$ value comes from the PTP1B system. All compounds in this set have two formal charges, and thus giving rise to further uncertainty when the

electrostatic contribution is evaluated. In general, ESMACS has shown a tight control over errors when binding affinities obtained from complex simulations only i.e. 1-trajectory ESMACS [41, 87, 43]. We have shown here that when ligands do not possess formal charges, ESMACS generates a low $\sigma_{average}$ value and thus gives rise to reproducible binding affinities.

4.3.2 CDK2 system: challenges involving sulphonamide parameterisation and ligand conformer selection

Closer visual inspection of ligand structures containing a sulphonamide group led to the discovery of unnatural dihedral angles. Th atoms involved in the diheral bond are C-C-S-N and a O-S-N for the bond anglel. The pre-simulation parameterised ligand structure showed a dihedral and bond angle of 89.8° and 105.6°, respectively, however, post-simulations the same angles were 53.8° and 87°, respectively. Upon assessing the force field parameters, we noticed that indeed, the General Amber Force Field (GAFF) parameters did not account for the sulphonamide group.

It is likely that the poor parameterisation is due to the partial atomic charges assigned for the N and S atoms in the sulphonamide group. This was confirmed after assessing the other three ligands that contain the sulphonamide group. Three inhibitors have an unmethylated nitrogen atom (L1S, LIU and L32), whereas one of them is dimethylated at the nitrogen (L29). The difference lies with the partial atomic charges. The partial charges assigned to N and S atoms in inhibitor L29 are -0.22 e and 0.92 e, respectively. In the remaining structures the partial charges for the same atoms are -0.94 e and 1.22 e, respectively. The difference in partial atomic charge between the N and S atoms is large in the unmethylated inhibitors, compared to inhibitor L29 and this difference causes the O-S-N bond angle to become acute, disrupting the conformation of the sulphonamide group, including the dihedral angle. Inhibitor L29, does not experience these changes, and this is supported with an expected bond angle of 107.28°, and dihedral angle of 96.75°.



Atom	GAFF		CGenFF
	L29	L1S	
C1	0.01	-0.23	0.24
S	0.92	1.22	0.61
O1	-0.53	-0.57	-0.42
O2	-0.53	-0.57	-0.42
N	-0.22	-0.93	-0.77
H1	–	0.41	0.38
H2	–	0.41	0.38
C2	-0.19	–	–
C3	-0.19	–	–

Figure 4.2: *The partial atomic charges generated by GAFF for the ligands L29 (top) and L1S (bottom) are reported. For comparison, the charges assigned by CGenFF are also included in the table. The structures on the left show the sulphonamide group. L29 contains two methyl groups substituted on the nitrogen atom where L1S contains only hydrogen atoms. The aromatic group belonging to atom C1 has been included, and the remainder of the ligand has been omitted.*

The unusual partial charge assignment was investigated further by comparing these values with another force field, namely CGenFF [68], developed by CHARMM. It was noticed that the partial charges, particularly on the N and S atoms are more representative of what was generated for L29 (Fig. 4.2). More specifically, when comparing L1S partial charges generated by GAFF, and CGenFF, the S atom is less positive, and the N atom less negative in the latter. It was also observed that the aromatic carbon C1 possess a negative charge, which is uncharacteristic for this atom type. A slightly positive charge on the C1 atom was achieved when using CGenFF, which is what is expected for this atom type.

Based on this, it was decided to complete an ESMACS run on a sulphonamide-containing ligand with CHARMM partial atomic charges. If the binding affinities

obtained from this model complement the successfully parameterised ligands, then it is very likely that the wayward results due to the sulphonamide-containing ligands is due to incorrect partial atomic charge assignment. This was done by constraining the partial charges for the sulphonamide group atoms (shown in Fig. 4.2), and generating an electrostatic potential, as completed previously. These new partial atomic charges were used and the ESMACS workflow was completed as before.

Indeed, Fig. 4.3 shows that the sulphonamide-containing ligands, with CHARMM partial charges, has integrated well with the remaining ligands. There has been a considerable improvement in ranking and correlation across all free energy methods, with a decent coefficient observed for the PB method. This highlights the importance of correctly parameterising ligands prior to MD simulations. Small changes in initial assignments plays a very sensitive role in the final outcome.

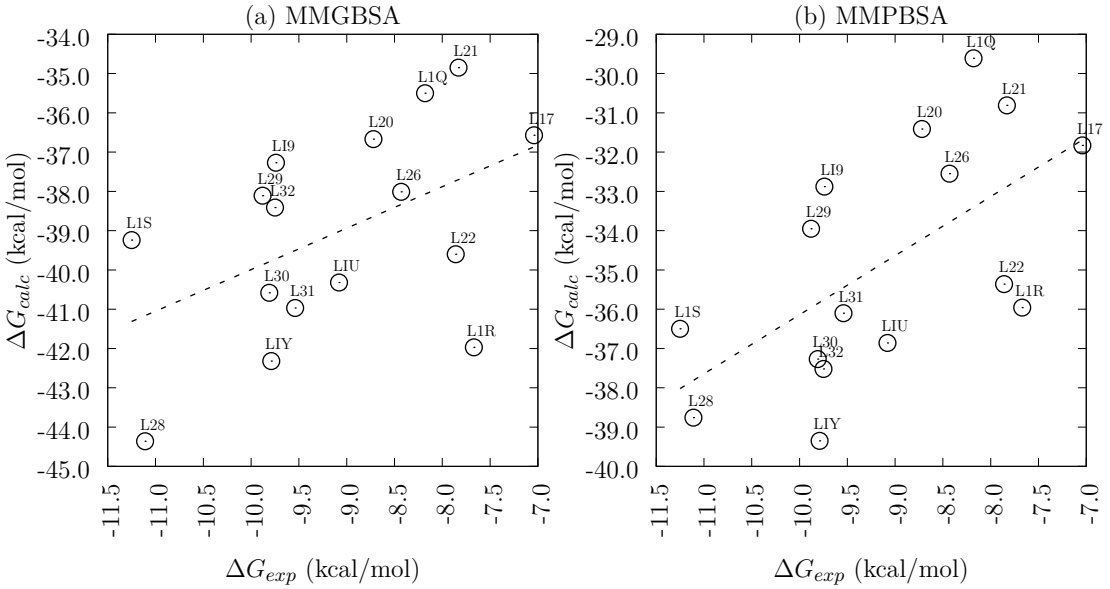


Figure 4.3: *Correlation plots for calculated and experimental ΔG values for 16 ligands complexed with CDK2, using the 1-trajectory ESMACS method.*

In this study, multiple rotamers are available for some of the ligands. It is important to find a method of selecting the correct rotamer conformation, rather than choosing the ligand that fits the correlation the best.

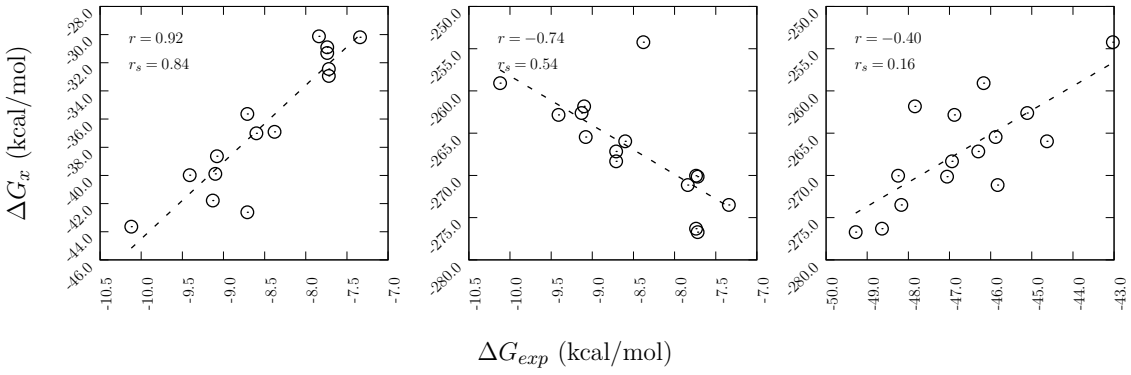
In some cases, the selection process is redundant: if a crystal structure is available than this can be used. In the CDK2 case, the starting crystal structure is ligand L1Q, which is essentially the sub-structure for all of the remaining ligands. Substitutions in the para position are allowed as this protrudes towards the bulk solvent and the chemical group has room for manouvre. Additionally, any movement around the rotable bond of the aromatic group will not change the arrangement in space. However, substitutions of chemical groups in the meta position mean that the aromatic group will rotate around the bond to find an energetically favourable arrangement.

The lowest energy ligand rotamer is selected by summing the free energy of the ligand and the calculated binding affinity ($G_{lig} + \Delta G_{calc}$). This not only takes into the account the lowest free energy of the rotamer, but also the interactions attributed to binding with the receptor, that is, the calculated binding affinity (ΔG_{calc}). Alternatively, the free energy of the complex ($G_{complex}$) can be assessed but this would introduce too much noise from all other interactions within the receptor and deem it unreliable.

We see that selecting rotamers based on the above criterion does not make any significant improvement in correlation. Equally, selection of the original data set (in other words, excluding the rotamer ligands), there is no improvement in correlation.

4.3.3 PTP1B system: aberrant electrostatic energy calculations result in a loss of correlation

The PTP1B system reports a negative correlation (Table 4.1). A closer look at the energetic terms that contribute to the binding affinity showed a very positive ranking and correlation for the non-polar terms (ΔG_{vdw} and ΔG_{surf}, and an opposite trend for the electrostatic terms (ΔG_{elec} and $\Delta G_{solv}^{PB/GB}$). Figure 4.4 shows the correlations of the above terms with experimental values.



(a) $x = vdw$ (b) $x = elec$ (c) $x = calc$

Figure 4.4: *Correlation plots of: (a) the van der Waals (ΔG_{vdw}) energy contributes to the binding affinity versus the experimental binding affinity (ΔG_{exp}), (b) the electrostatic contribution (ΔG_{elec}) to the binding affinity against ΔG_{elec}, and (c) the final binding affinity (ΔG_{calc}) versus ΔG_{exp}. All values reported are obtained via the PB free energy method.*

There are two probable reasons for this. We have already elucidated that the ligands in the PTP1B set all contain two formal charges (both are deprotonated carboxylic acid groups). Formal partial charges have generally received mixed ranking correlations [132], using the MMPBSA and MMGBSA approach. This is likely the case here. The second issue is that the binding pose of this data set is such that two crystal water molecules are trapped within the pocket. Using the implicit solvent models, the electrostatic contribution associated with presence of these two crystal water molecules are disregarded.

Table 4.2: *A table describing summarising the ligand transformations, along with the corresponding relative free energies. All free energy values are reported in kcal/mol.*

Transformation	Initial Ligand	Final Ligand	$\Delta\Delta G_{exp}$	$\Delta\Delta\Delta G_{TIES}$	$\Delta\Delta G_{vdw}$
T01	L73	L68	-0.64	-0.14	-6.01
T02	L66	L74	1.26	1.19	9.77
T03	L74	L72	-0.50	0.04	-2.88
T04	L74	L77	1.01	0.80	3.75
T05	L78	L77	1.04	1.54	5.00
T06	L75	L76	0.69	0.68	4.34
T07	L80	L79	0.42	-0.61	-0.84

As a result of the aforesaid, the $\sigma_{average}$ for the electrostatic contribution is approximately 10 kcal/mol. This is the main reason for the high $\sigma_{average}$ seen in Table. 4.1. On the other hand, the $\sigma_{average}$ for the non-polar van der Waals contribution is around 2 kcal/mol. We also see a good correlation and ranking for the van der Waals contribution, shown in Fig. 4.4a. In the PTP1B case, only the non-polar contributions are required to generate strong correlations, and this could be the case for other ligand sets with high formal charges.

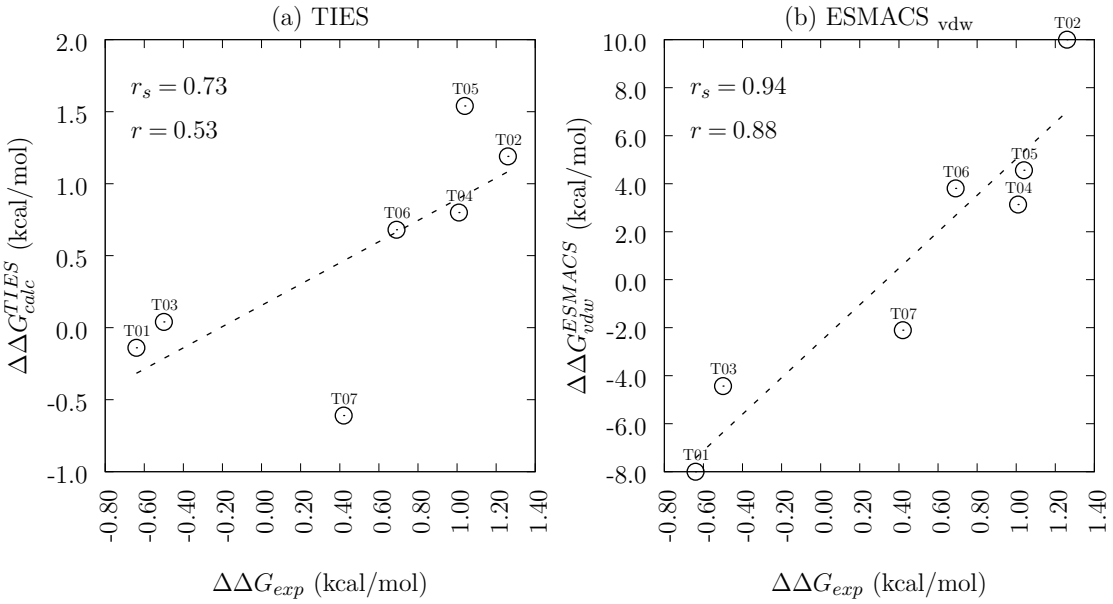


Figure 4.5: *A correlation plot of: (a) relative binding affinities between 7 ligand pairs, obtained using the TIES protocol ($\Delta\Delta G^{TIES}_{calc}$), and (b) relative van der Waals energy of the TIES ligand pairs, obtained using the 1-trajectory ESMACS approach (ΔG $DeltaG^{ESMACS}_{vdw}$). We see a better ranking and correlation when considering $\Delta\Delta G^{ESMACS}_{vdw}$, but $\Delta\Delta G^{TIES}_{calc}$ produces more accurate results. Table 4.2 explains the ligand pairs used in this analysis.*

Taking the impressive correlations from the van der Waals term further, we compared this with TIES results. Fig. 4.5 shows the ranking $\Delta\Delta G^{TIES}_{calc}$ value obtained for some TIES transformations, and the change in $\Delta\Delta G_{vdw}$ between corresponding ligands that was obtained using ESMACS. We report that the latter, $\Delta\Delta G^{ESMACS}_{vdw}$ reports better rankings and correlation than the original TIES results. With this said, TIES still generates accurate binding affinities, where ESMACS results are

precise, and thus generate only good rankings.

4.3.4 ESMACS distinguishes between TYK2 ligand chemical groups

ESMACS binding affinities for the TYK2 system, using only the MMPBSA and MMGBSA free energy method (as opposed to including configurational entropy contributions using normal mode analysis), shows no ranking and minimal linear correlation (Fig. 4.6a). However, the inclusion of configurational entropy allows us to distinguish between chemical signatures within the ligand set, shown in Fig. 4.6b. The TYK2 ligand set comprises two sub-sets, that is ligands that contained aromatic and aliphatic variable groups. The total r_p and r_s is 0.17 and 0.03 respectively, but if we split the data set into the respective ligand sub-sets, we see that the aromatic and aliphatic sub-set perform considerably better when treated separately. We see rankings and correlations of 0.86 and 0.74, respectively.

In this instance we report a promising result using normal mode analysis (NMA) to obtain the configurational entropy term which contributes to the total binding affinity. However, NMA has traditionally received mixed results and its use is greeted with skepticism, due to the large compute requirements, and approximate nature of generating entropy estimates. This is depicted in the case of CDK2 where we also applied NMA, and as a result, we see a degradation in ranking and correlation.

4.3.5 Thrombin and MCL1 systems give rise to good correlation and ranking metrics

ESMACS generated good binding affinities for the thrombin and MCL1 systems, all reporting ranking and correlation coefficients above 0.60 and 0.80 respectively. Thrombin performed particularly well, generating an r_p of 0.92 and r_s of 0.84 for

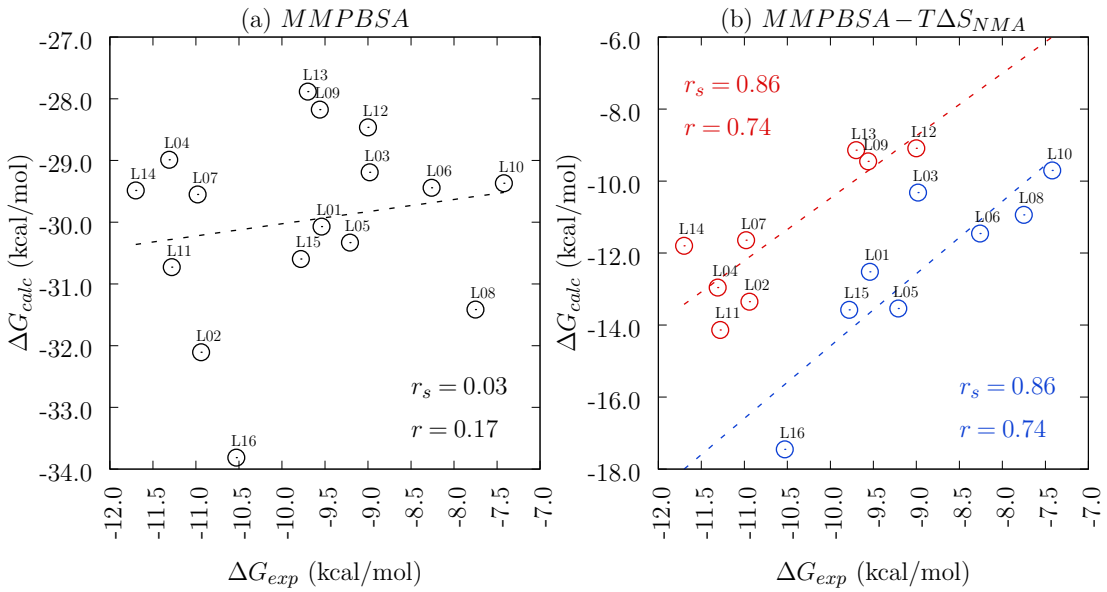


Figure 4.6: *A correlation plot of: (a) the binding affinities of 16 TYK2 ligands using the MMPBSA method, and (b) the MMPBSA free energy method is coupled with configurational entropy estimates, generated by normal mode analysis (S_{NMA}). The inclusion of the entropy term considerably improves the ranking for two ligand sub-sets that display specific chemical signatures, and thus allow us to distinguish between different chemical groups within a data set.*

PB approach.

We see that good binding affinity predictions are generated when studying smaller systems of study. Both thrombin and MCL1 are the smallest systems here, and it is likely that we obtain good results for this reason.

4.3.6 TIES as a tool in drug discovery

TIES results correlate very well with experimental binding affinities for the systems that have been selected in this chapter [42]. Two statistical measures were used to determine accuracy, which are the root-mean-square error (RMSE) and mean absolute error (MAE). Both are measure of the differences in values between calculated and experimental free energies of binding. The only difference being MAE does not take direction into account, as all values are absolute. Both RMSE

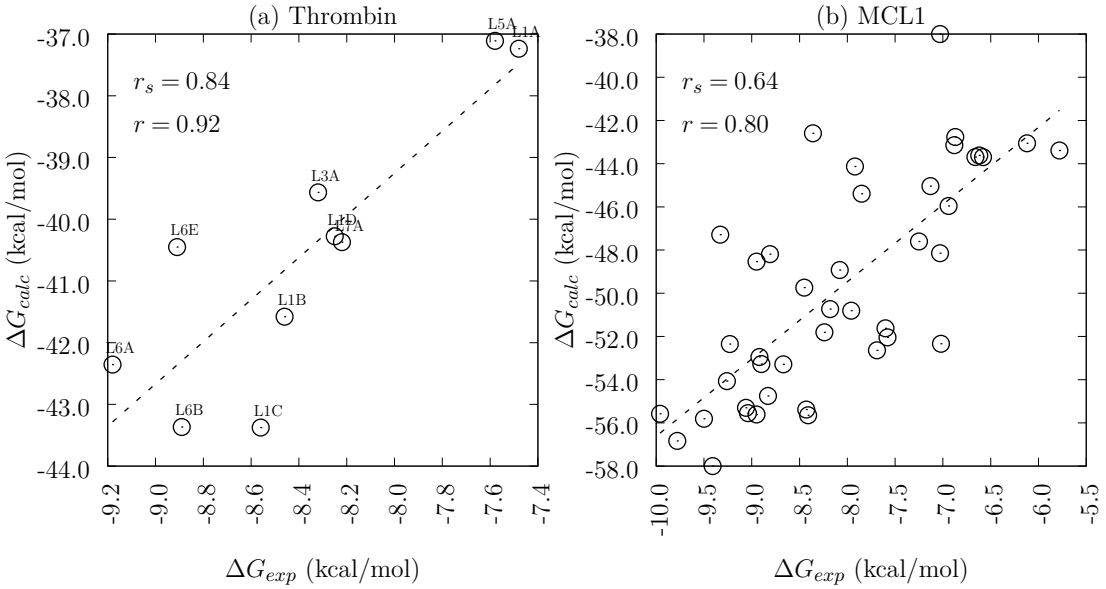


Figure 4.7: *Correlation plots using the ESMACS 1-trajectory approach with MMPBSA. Fig. (a) reports 10 thrombin binding affinities where as Fig. (b) shows binding affinties for 42 MCL1 ligands.*

and MAE indicate very high levels of accuracy. With regard to r_p and r_s, this too, performed well with coefficients values that represent very strong linear dependence and ranking, respectively. Reproducibility is defined by the control of errors. TIES has shown to keep a very tight control of errors, this is depicted in Fig. 4.8. A more detailed description of TIES performance on these systems has been published recently [42].

There were, however some difficulties with ligands that contained a sulphonamide group (L1S, LIU and L32). Although L32 produced a satisfactory $\Delta\Delta G_{calc}$ value, L1S and LIU calculations were aberrant. Inclusion of these three binding affinities drastically degrades the statistical measure that is employed. This is attributed to challenges in parameterising the ligand which was also experienced in the ESMACS study. We incorporated the same parameters (Fig. 4.2) in TIES, as was used in ESMACS, but this did not fix this issue.

4.3.7 Can we obtain equally good TIES results using fewer λ windows?

TIES calculations, using the default settings, require a large amount of compute resources. It would be of considerable interest if the core count could be decreased without sacrificing accuracy and precision. Simulation length, and replica number per λ window are set at the lower limit and so further decreases in these parameters would most likely mean insufficient phase space sampling [42], per λ window. Selecting which λ states to sample could allow for more compute efficient calculations. We applied this to the systems described here.

We first tested the the role of sampling a selection of λ windows, ranging from 3 to 11 for all transformations, and subsequently generating $\Delta\Delta G_{calc}$ values with standard errors. Then we compile the performance metrics (RMSE, MAE, r_p and r_s) used previously. We also included the original set which we used to compare our results.

Intuitively, as the number of λ windows sampled increases, we generally see an improvement in all metrics. Particularly, when increasing from 3 to 5 windows. However, in the case of TYK2 and MCL1, there is not a significant deviation in metrics between 5 windows and the original full set of 13 windows. This would suggest that, at least initially for these systems, it would have sufficed to run 5 windows. This corresponds to nearly a 30% decrease in compute resources.

However, we also see a trend in all systems which is not entirely expected. In some instances, when λ windows are increased, a degradation is reported in the performance metrics. This was assessed further by performing λ 'exclusions' (Fig. 4.9).

$\Delta\Delta G_{calc}$ with standard errors were generated, but with the omission of selected λ states. This was completed serially for each λ window. In other words, we obtained 13 sets of performance metrics, and in each, one λ window has been excluded. This would give us some indication of how each λ window is performing, and why we

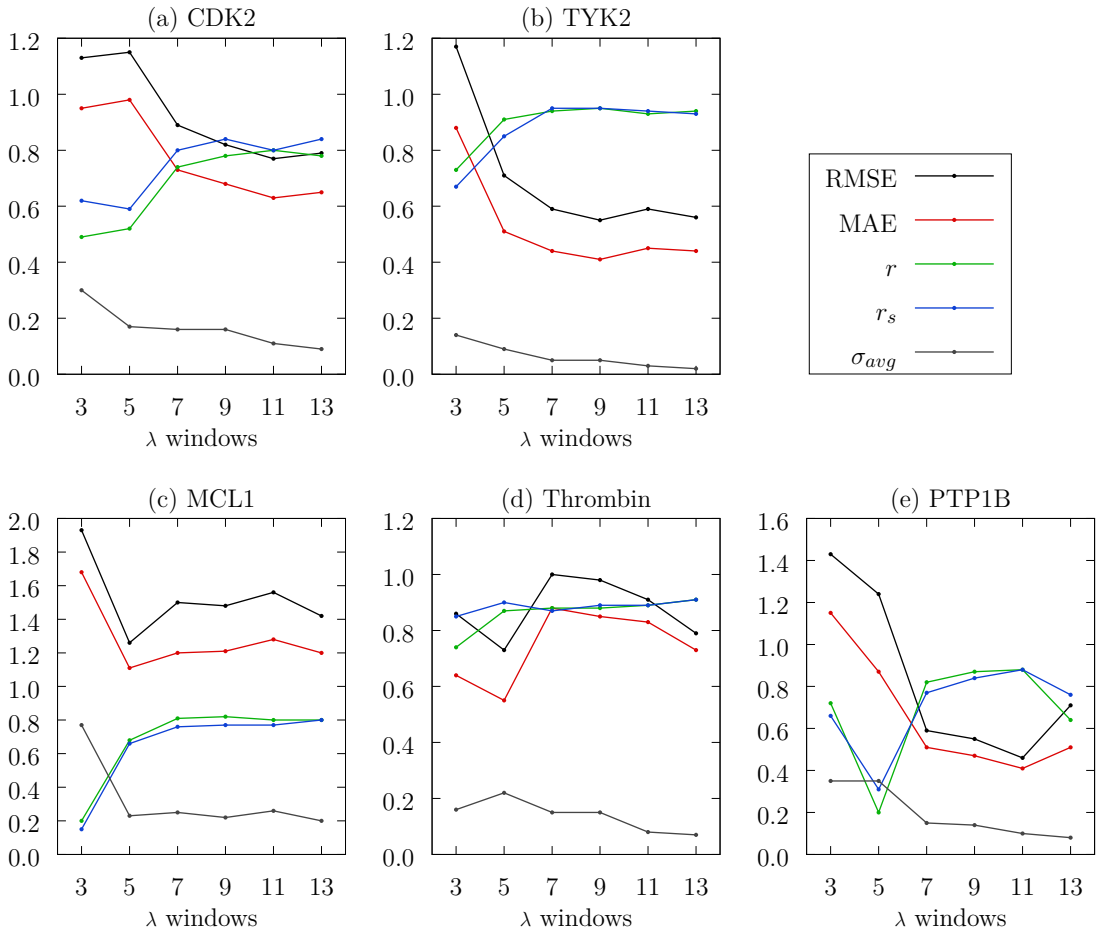


Figure 4.8: *A plot of correlation and prediction metrics for TIES relative binding affinities, extracted from a various λ window selections. Plots are shown for each system of study, and each line represents a different metric with its assignment shown in the key. The axis labelled 'λ window' is the amount of λ within the selection. For example selection 1 has 3 λ windows. 13 λ windows is the original data set, with relative binding affinities obtained from all λ windows.*

see a worsening in metrics, when λ windows are increased.

The PTP1B system shows a large amount of fluctuation when different λ windows are excluded. A likely reason for this is the formal charge that exists in the PTP1B ligands. Although there is a large fluctuation in this system, the σ_{avg} value is stable and low, suggesting that sufficient sampling has been completed. All other systems shows significant fluctuations as each λ window is excluded. In all systems, we see

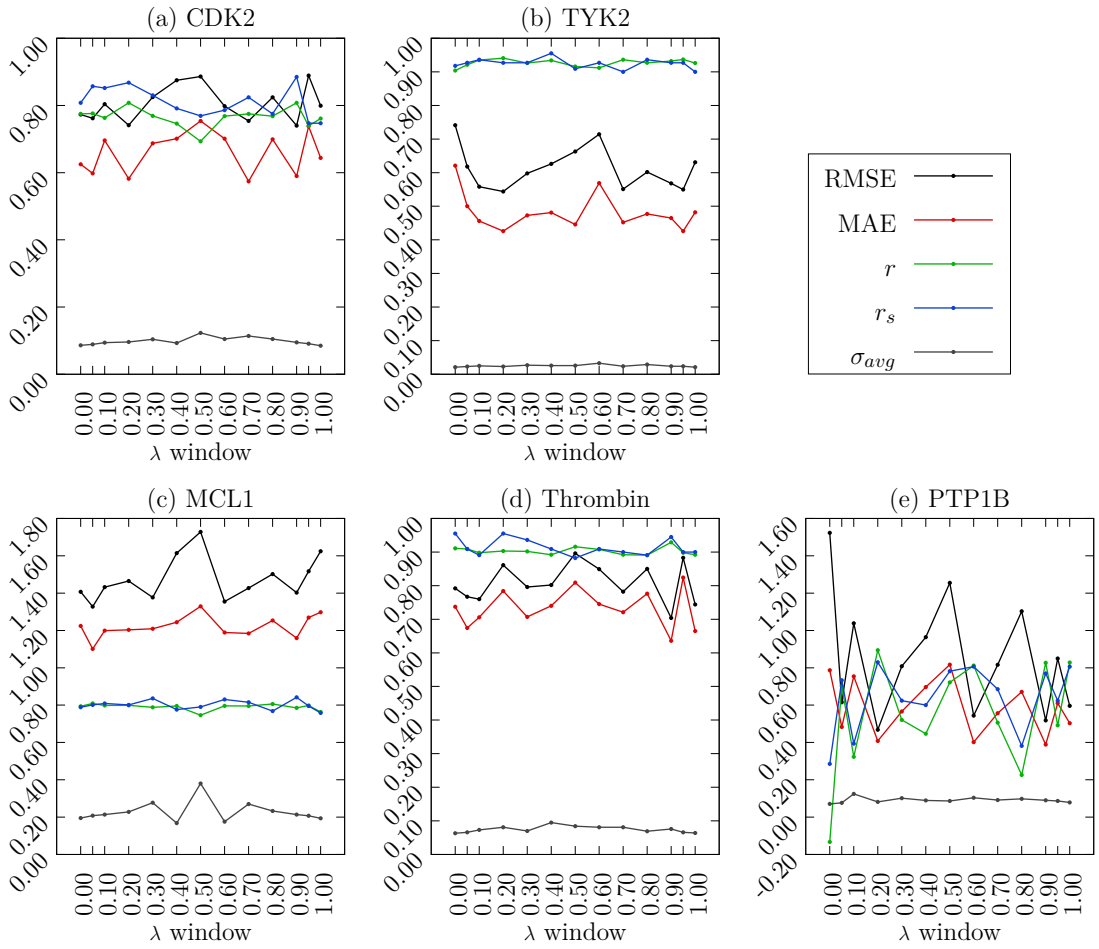


Figure 4.9: *A plot of correlation and prediction metrics for TIES relative binding affinities, extracted from a various λ window selections. Plots are shown for each system of study, and each line represents a different metric with its assignment shown in the key. The λ sets are selected by excluding one λ window in serial fashion. Thus the x-axis label 'λ window' is the λ that has been excluded. For example λ window 0.00 means that this window has been excluded and all others have been used to generate relative binding affinities.*

a peak at $\lambda = 0.50$ (and surrounding λ windows) which corresponds to a worsening of performance metrics. Similarly, we see peaks at the end points, between 0.00 and 0.10, and 0.90 and 1.00. This indicates that near the end points, a higher resolution of λ windows is required. This applies also for the region between 0.40 and 0.60.

4.4 Discussion

ESMACS generates good correlations for 5 systems of study, using the MMPBSA free energy method. The MMGBSA method consistently performs worse compared to MMPBSA. Additionally, the 1-trajectory approach is beneficial in all cases. The $\sigma_{average}$ for all systems is very low – no higher than 0.48 kcal/mol – indicating that ESMACS is able to generate reproducible and precise binding affinities. We see that in the case of CDK2, that the correct parameterisation of ligands is vital. Here, the sulphonamide group had to be reparameterised manually, which gave improved results. ESMACS, or TIES, does not take into account different ligand conformations. This means one conformation has to be selected that is believed to be correct, which is not a theoretically rigorous approach that can be used in drug discovery programmes.

We see in the PTP1B system that the electrostatic contribution was incorrectly generated resulting in a negative correlation when correlated with experimental binding affinities. Conversely, the non-polar terms performed impressively, and even produced a better ranking than TIES when correlated with relative binding affinities. Care must be taken when completing ESMACS calculations with ligands that contain high formal charges.

In the TYK2 system, we are able to distinguish between chemical signatures when we include entropy calculations. ESMACS binding affinities with MMPBSA alone showed modest correlation, but when entropy values were included, we see a good correlation and ranking for two different chemical groups. Thus ESMACS, with the use of entropy estimates, could allow us to distinguish between subsets of a chemical data set, and subsequently generate good correlations. Both MCL1 and thrombin perform well using the MMPBSA approach.

TIES results show good accuracy as portrayed by the RMSE and MAE metrics. As a result it also generates good rankings and correlations [42]. Standard errors are

extremely low, indicating reproducibility. Parameterisation was also an issue here, for the CDK2 ligands that contained sulphonamide groups. However, the modified parameters did not perform as expected in TIES, as it did in ESMACS.

Running TIES calculations requires a large amount of compute resources and so reducing the CPU usage without compensating on performance would be desired. In the TYK2 and MCL1 systems, running only 5 λ windows has shown similar performance metrics to running all 13 windows. In CDK2 and thrombin, running 7 λ windows will generate comparable metrics to completing all windows. Therefore, a decrease in λ windows is a plausible approach to save on compute resources and still achieve acceptable results.

When excluding λ windows serially, we see that there is a large fluctuation in performance metrics. This is particularly the case at end points, and at λ windows surrounding 0.50. It is likely that further λ windows should be included to improve TIES performance.

In conclusion, we have shown that the both ESMACS and TIES generate reproducible binding affinities across 5 different proteins and 99 ligands varying in size, flexibility, charge and biological function. The reproducibility is characterised by tight control of error bars through the use of replica simulations. TIES binding affinities are shown to be accurate whereas ESMACS generated precise free energy values. The current limitations of ESMACS and TIES is that the user is required to select the correct ligand conformation prior to simulation requiring a highly technical user and additional investigation of the correct ligand conformation. Both protocols are limited by the parameterisation of the ligands. For example, ligands containing the sulphonamide group required reparameterisation which resulted in binding affinities with an improved correlation and ranking coefficient.

Chapter 5

Towards improved solvent models for the prediction of binding free energy and configurational entropy

5.1 Introduction

A major challenge in computational drug design is developing realistic solvent models to be able to generate accurate binding affinities. Some approaches, such as TI and FEP, are accurate and capture solvent effects as water molecules are treated explicitly. These 'exact' free energy methods are computationally more expensive than 'approximate' methods, namely MMPB/GBSA, and thus require a large amount of CPU time to estimate the binding affinity of one receptor-ligand complex. TI/FEP are often limited to smaller and chemically related data sets. MMPB/GBSA is thought to be good balance between accuracy and compute requirements when estimating binding affinities. Section 2.6.1 and 2.6.2.3 explains how binding affinities are obtained from from TI/FEP and MMPBSA, respectively.

In this chapter we will work towards improving the current models by understanding the relationship between atomic partial charge assignment, and subsequent calculation of electrostatic energy terms. This can be done by optimising current implicit solvent models. We also employ a solvent accessible surface area-based entropy method and assess its performance compared to normal mode analysis.

5.1.1 Estimating the free energy of binding

MMPB/GBSA method applies the approximation of a continuous medium for the solvent that describes the internal and external dielectric constant of the protein and water, respectively. This is known as a continuum or implicit solvent model. Applying a continuous medium requires the polar contribution to the free energy of solvation, ΔG^{sol}_{pol}, to be solved using the Poisson-Boltzmann (PB) or Generalised Born (GB) formula. Although implicit solvent models reduce compute cost and minimise background solvent-solvent interactions, important solvent effects are often over-looked which would otherwise make significant contributions to the binding free energy.

Much work has been done to improve the prediction of binding free energies by improving the solvent models used. One alternative is to simply add a small number of explicit water molecules (or even a single water molecule) that play a direct role in ligand binding. Here, important solvent-solvent and solvent-solute interactions are preserved whereas the remaining solvent is modelled implicitly. Inclusion of a handful of water molecules that play an important role in ligand binding will not result in excessively large fluctuations in energy. Studies in this domain have generated mixed results. For instance, Greenidge et al. showed that the inclusion of select explicit water molecules in fact slightly worsened the predictive performance of MMGBSA on a large and diverse set of ligands [133]. Equally, Checa et al. [134] have shown that the inclusion of explicit waters results in poor correlation coefficients when studying trypsin inhibitors. On the contrary, Wanoefer et al. [135]

and Wan et al. [100] show improved correlations when a single, and a network of, crystal waters were included, respectively. Only one of these studies [100] executed multiple MD simulations from which reproducible binding affinities were obtained. Binding affinities obtained from single MD simulations are not reliable.

Alternative methods include adjusting the internal dielectric constant (ϵ_{int}) before free energy calculation as the ϵ_{int} is more difficult to predict due to the polarity of residues that occupy a binding pocket. Modifications of ϵ_{int} have been shown to be very sensitive, with varying degrees of change in correlation coefficients with experimental values. Hou et al. [136, 137] have shown that ϵ_{int} should be system dependent; increasing the ϵ_{int} for the α-thrombin system yields an improvement in ranking, where the same increase in ϵ_{int} produces a degradation in ranking of other systems. Genheden and colleagues [138] have noticed that ϵ_{int} values between 1 and 25 all show largely varying correlations. Genheden et al. [139] recognise the importance of using an ensemble of short, independent MD simulations to produce statistically converged binding affinities using the MMGB/PBSA free energy method.

Fundamentally, there have been mixed results on performance when incorporating explicit water molecules into MMPB/GBSA calculations, and modifying the ϵ_{int} parameter, and there is yet to be a universally accepted method that assigns correct explicit water molecules or ϵ_{int}, respectively, to any given biological system, which consequently achieves consistently improved binding affinity predictions.

The quality of binding affinity predictions are highly dependent on the type of molecular mechanics force field that is selected. In the AMBER force field that is implemented in ESMACS and TIES, the molecular mechanic partial charges is fitted to reproduce the quantum mechanical electrostatic potential of small molecules which is a good strategy to replicate multipole moments and electrostatic interactions. The disadvantage is that this type of force field fails to take into account polarisation by the varying dielectrics of the environment. The incorporation of a

polarisable force field in ESMACS/TIES is a legitimate avenue of exploration. The development of polarisable force fields is a fruitful area of research with a number of models available. A recent review [140] highlights the development of robust polarisable force fields (AMBER ff02, AMOEBA, Drude) that are compatible with existing, popular models.

5.1.2 Estimating the configurational entropy of binding

The S^X_{conf} term in Eq. 2.63 is the configurational entropy associated with ligand-receptor binding, and is the sum of the entropies of the vibrational, translation and rotational degrees of freedom. This term is often estimated using normal mode analysis (NMA), employing the rigid-rotor harmonic oscillator (RRHO) approximation.

NMA is a technique used to assess the vibrational degree of freedom of a harmonic oscillating system, which is at, or very close to, its equilibrium [141]. Computationally, this requires minimisation of a simulation snapshot, building a Hessian matrix, and subseqeuntly diagonalising the matrix to gain frequencies. Hessian diagonalisation scales as $(3N)^3$ where N is the number of atoms in the system [142]. These calculations are computationally expensive, and require large-memory cores, which are not always available. Due to compute and memory requirements, NMA is usually performed in a vacuum and on a truncated system with approximately a 10 Å radius around the protein-ligand binding region. As a result of these changes to the original simulation snapshot, values from the normal mode method are not an accurate representation of the simulated system, and so the configurational entropy is largely approximate.

As a result, a solvent accessible surface area-based method [86] was employed to estimate configurational entropy term in Eq. 2.62. Note that Wang et al. refer to the S^X_{conf} term as 'conformational' entropy, but here we use 'configurational' entropy. One of the main strengths of this method is that configurational entropy

estimates can be generated on a commercial PC within minutes. In addition, estimates are not based on a minimised structure as in NMA.

A macromolecule is made up of atoms that are exposed to solvent to different extents. Interior atoms have more restricted movement and thus contribute less to entropy, however, it is incorrect to assume that these atoms play no part in entropy changes associated with ligand binding. In other words, atoms that are completely buried play a non-negligible role in entropy and must be included.

$$S_{WSASA} = \sum_{i=1}^{N} w_i(\mathrm{SASA}_i + k\mathrm{BSASA}_i) \tag{5.1}$$

In Eq. 5.1, w_i is the weight of atom i; SAS_i and BSASA_i are the solvent and and buried solvent accessible surface areas of atom i, and N is the number of atoms in the model system . The term BSASA_i is estimated using Eq. 5.2 where r_{prob} is the probe radius which has been set to 0.8 Å .

$$\mathrm{BSASA}_i = 4\pi(r_i + r_{prob})^2 - \mathrm{SASA}_i \tag{5.2}$$

The term k is an adjustable parameters that allows the user to adjust the extent to which the buried atoms (BSASA) contribute to S. Setting the k parameter to 0 means that the buried atoms do not contribute to the entropy, and conversely, setting k to 1 means that the buried and solvent accessible atoms contribute equally to the estimate of S. Estimating entropy using this approach is termed 'WSASA'.

5.1.3 Role of PAK4, BACE1 and ROS1 in pathogenesis

Below is a brief explanation of how each system is implicated in disease, giving context to the inclusion in this chapter.

PAK4

P21 protein-activating kinases (PAKs) are orchestral in regulating the formation and organisation of the actin cytoskeleton in mammalian cells, which induces changes in cell growth, morphogenesis, adhesion and proliferation [143]. PAKs fall into two categories: Group A PAKs, consisting of PAK1, 2 and 3. This group of PAKs may not have a direct role in cytoskeletal organisation [144], and thus their roles are not entirely understood. Group B PAKs, consisting of PAK4, 5 and 6. PAK4 is expressed ubiquitously [145], where both PAK5 and 6 are predominantly expressed in the brain, amongst other tissues [146, 147, 148].

Group B PAKs have highly conserved binding domain and kinase domain sequence structures, which hold a 50 percent homology with group A PAKs. Outside of the binding and kinase domain, Group B PAKs have no resemblance to each other, or that of Group A PAKs [144].

PAK4 acts as an effector, a protein that selectively binds to its target and regulates biological function, of the cell division control protein 42 homolog (CDC42) protein [149, 150]. CDC42 is a GTPase, which is a large family of enzymes that hydrolyse guanine triphosphate (GTP) into guanine diphosphate (GDP). CDC42 is part of the Rho family of GTPases that have been linked with cytoskelton dynamics and organelle development and other important cellular functions [145]. PAK4 interacts with the active conformation of CDC42, through the GTPase-binding domain. Subsequently, the PAK4-CDC42 complex induces redistribution of PAK4 to the Golgi membrane which triggers actin formation necessary for cell growth and proliferation. The formation and reorganisation of the actin cytoskeleton is dependent on PAK4 activity and its interaction with CDC42. Over-expression of PAK4 sees the genesis and profileration of highly-disorganised cancer cells in various tissue types [151, 152].

BACE1

Over the past 2 decades, the role of BACE1 [153, 154, 155, 156, 157] has been linked with the development of Alzheimer's Disease (AD). Therefore, developing an BACE1 inihibitor is an ongoing area of research with the goal of treating AD.

Pathogenesis of AD, in part, is a result of the accumulation of plaques, or peptides, in the brain, known as β-amyloid (Aβ). The formation of Aβ is a sequential catalytic process beginning with the cleavage of the amyloid precursor protein (APP), by β-secretase enzyme, into the N-terminus Aβ, and the C-terminal fragment C99. Following this, the γ-secretase enzyme cleaves C99, which gives rise to additional Aβ. The γ-secretase enzyme splices the C99 peptide in varying lengths and so different isoforms of Aβ are present. It is the longer isoforms that are responsible for the onset of AD.

Research has shown that there are over 200 mutations in the genes that code for APP and γ-secretase enzyme [158], which are directly associated with familial AD (FAD). In fact, the most widely known mutations (K670N and M671L) [159], are found close to the catalytic side, of β- and γ-secretase, and hence support the cleavage of APP. This causes a larger deposit of Aβ isoforms in cerebral tissue. Conversely, the A673V mutation in APP results in less effective cleavage by β-secretase, and so Aβ is decreased by 40% [160].

The β-secretase enzyme was identified as a key target for the treating FAD, and was subsequently renamed 'β-site APP cleaving enzyme' (BACE1) [153]. BACE1 is an aspartic acid protease that is 501 residues long, with a signature sequence of DTGS and DSGT that characterise the aspartic acid proteolytic site [161].

ROS1

ROS1 is a serine/threonine kinase that plays a role in oncogenesis. Little is known about the function of wild-type ROS1. However, mutant forms of ROS1 create

fusion proteins, in which the catalytic region of the kinase becomes constitutively active [162]. As a result, this drives uncontrolled cellular proliferation. ROS1 fusion proteins have most recently been linked non-small cell lung cancer (NSCLC, [163]).

5.1.4 Motivation

In this chapter, we take a PAK4 data set that shows little correlation with experimental results, and attempt to improve the solvent model by: (a) including explicit water molecules that may make a considerable contribution to the binding affinity and (b) modifying the ϵ_{int} which has shown to improve binding affinity predictions in some systems. In the same system, we witness a degradation of correlation when the configurational entropy calculated using NMA, is included. Thus, we employ the WSASA entropy method and assess its performance. The above was extended to other challenging systems: BACE1 and ROS1.

The PAK4 receptor and corresponding ligands were extracted from a publication by Crawford et al. [164]. The rationale for selecting this system of study was primarily due to the ligands in question; structurally there are two sub-sets which contain an aliphatic and aromatic ring structure in the variable group. We were interest to see how ESMACS performs across ligands with different ring structures, within the same binding domain. A distinct difference in the performance of ESMACS was seen in the TYK2 system described in chapter 4 and it is possible that a similar trend could be witnessed here. BACE1 and ROS1 were subsequently included for the same reasons. BACE 1 represents a ligand data set with a ring structure of varying size. The ROS1 ligand data set, like PAK4 ligands, contain saturated and unsaturated rings.

5.2 Methods

ESMACS, which includes the 1-, 2- and 3-trajectory variants, were employed for all systems in the study. PAK4 binding affinities were generated using two different

MD trajectories: one group of simulations included crystal water molecules whereas the other group disregarded crystal waters. Preparation of these models are described here. Binding affinity calculations involving changes in ϵ_{int}, and inclusion of explicit water molecules, are also described in more detail, here.

5.2.1 Model preparation

Complex models that have been used to complete MD simulations can be divided into two sub-sets, each containing 14 identical ligand-receptor complexes. One contains crystal waters present in the PDB file, where the other set has been stripped of crystal water molecules, from this point, both complexes were then solvated using *tleap*. The PDB code for the receptor without crystal waters is 5BMS [164], and with crystal waters is 2X4Z [165]. For the simulations with retained crystal waters, all 142 molecules within the PDB file were kept.

Fourteen ligands (Fig. 5.3) were selected, which originate from an experimental study [164]. Ligand L01 is a previous pre-clinical drug candidate which failed on the grounds of poor oral bioavailability [165]. All remaining ligands were created by extracting the skeletal structure from PDB codes 4ZY4, 4ZY5 [164], and the variable moiety manually modified using Schrödinger's Maestro Suite (free of charge). Ligands L01 and L04 came directly from PDB codes 4ZY4 and 2X4Z, respectively. These ligands were then docked into the respective PAK4 receptor models described above.

With regard to the experimental binding affinity of L01, the binding affinity from the initial study [165] was used, whilst the remaining values were extracted from Crawford et al. [164]. Error bars are not available for any of the experimental ligand binding affinities. A more thorough review on experimental binding affinities is presented in section 2.5.3.

The initial model of BACE1 was based on PDB 3ZOV [166] with loops modelled

by Janssen. Binding poses were provided for 21 diverse ligands as shown in the Fig. A.7. BACE1 was assayed fluorometrically on Monaco Safas spektroflourometer [167].

The initial model of ROS1 was based on an X-ray structure determined in-house by Janssen. Missing sequence Lys2117–Gly2121 has been extracted from a non-disclosed in-house ROS1 X-ray structure and merged in the structure. No local minimization has been performed on the rebuilt structure. The provided dataset contained 32 ligands (see Fig. A.6). The ligands have been manually aligned using compound JNJ-54192398 from the X-ray structure as a reference. Alignment was performed with the Template CoMFA flexible alignment tool.

5.2.2 Explicit water free energy calculations

The protocol applied here has been adapted from a previously published method [168]. The same frames that were used for the implicit solvent calculations were also used here. The closest water molecules within n Å of the ligand, were selected for each frame using the *closest* flag within *cpptraj* package, developed by AMBER[89]. Seven sets of explicit water calculations were completed with varying amounts of water molecules in the vicinity of the respective ligands. The values of n were 10 to 70, increasing in increments of 10. The water molecules were included as a part of the receptor and the remaining solvent was removed. Binding free energies were calculated using the MMPB/GBSA models based on same protocol as previously described.

5.2.3 MMPB / GBSA calculations with varying internal dielectric constants

The ϵ_{int} parameter can be adjusted in the MMPBSA module, however a more temporally efficient approach is to simply divide the electrostatic component of the G^{MM}_{elec}, and the G^{X}_{sol} in Eq. 2.52 and 2.53, respectively, by the ϵ_{int} value, after

completing MMPBSA calculations with $\epsilon_{int} = 1$. This yields the same results as completing the MMPBSA calculation with an adjusted ϵ_{int} parameter, and has been reported in a previous study [136]. The external dielectric constant, ϵ_{ext}, has been maintained at 80 throughout all calculations.

5.3 Results

The respective correlation and ranking metrics, Spearman (r_s) and Pearson (r_p) coefficients, perform poorly for binding affinity predictions with and without crystal water molecules, in the 1-, 2- and 3-trajectory approach. There is a slight improvement when crystal waters are introduced but this is negligible. It is also evident that there is no significant difference in correlation and ranking between binding affinities, regardless of the inclusion of configurational entropy, in either the GB or PB method.

Within these results, we obverse some interesting trends. Firstly, ligand L01 has a much larger negative ΔG_{bind} value than expected. Additionally, the inclusion of crystal water molecules in this model causes the largest change in ΔG_{bind} than any other ligand. The binding affinity becomes more positive by approximately 7 kcal/mol in all free energy methods. The importance of crystal waters when calculating binding affinities is seen here: it is evident that specific water molecules contribute heavily to binding, and disregarding this results in aberrant calculated free energies. Secondly, any semblance of correlation and ranking that has been reported thus far, is responsible for the large negative ΔG_{bind} attributed to ligand L01 (Table B.11). Removing this data point, however, results in the absence of a correlation or ranking shown in Fig. 5.1. This lack of correlation was the motivation for the study. Therefore, two investigations will take place: the first is to identify the reason for the large fluctuation in ΔG_{bind} when crystal water molecules are included in the PAK4-L01 complex; and second is to understand why the PAK4 complexes studied here seem to insensitive to ESMACS.

Table 5.1: *Binding free energies using the 1-trajectory ESMACS approach for PAK4 ligands bound to the PAK4 receptor when crystal water molecules were included. GB/PB is the free energy method using the Generalised Born or Poisson Boltzmann approximation. GB/PB_{NM} is the above method with the inclusion of configurational entropy term. All values are in kcal/mol.*

	1-trajectory				
Ligand	GB	PB	GB_{NM}	PB_{NM}	ΔG_{exp}
L01	-52.58	-39.17	-29.60	-16.18	-11.70
L03	-41.17	-31.89	-19.43	-10.15	-11.66
L04	-39.98	-32.18	-17.47	-9.69	-10.62
L18	-36.84	-30.37	-14.94	-8.47	-11.26
L19	-40.20	-29.91	-17.47	-7.24	-8.38
L20	-39.42	-31.53	-15.21	-7.31	-8.33
L22	-39.99	-32.15	-17.45	-9.61	-9.17
L23	-40.83	-33.22	-19.87	-12.27	-9.89
L24	-40.35	-31.39	-18.43	-9.47	-8.07
L25	-39.99	-32.35	-21.76	-14.13	-7.75
L27	-40.06	-33.56	-26.20	-19.70	-9.35
L29	-42.24	-32.81	-32.60	-23.17	-8.50
L30	-43.36	-32.28	-22.58	-11.51	-8.80
L31	-45.89	-33.71	-44.39	-32.21	-8.39

5.3.1 Crystal waters within the binding pocket of L01 play a critical role in binding affinity predictions

We begin with the investigation of L01 and the considerable contribution crystal water molecules have to binding. To understand what causes this, the trajectories of simulation states, with and without crystal water molecules, were analysed.

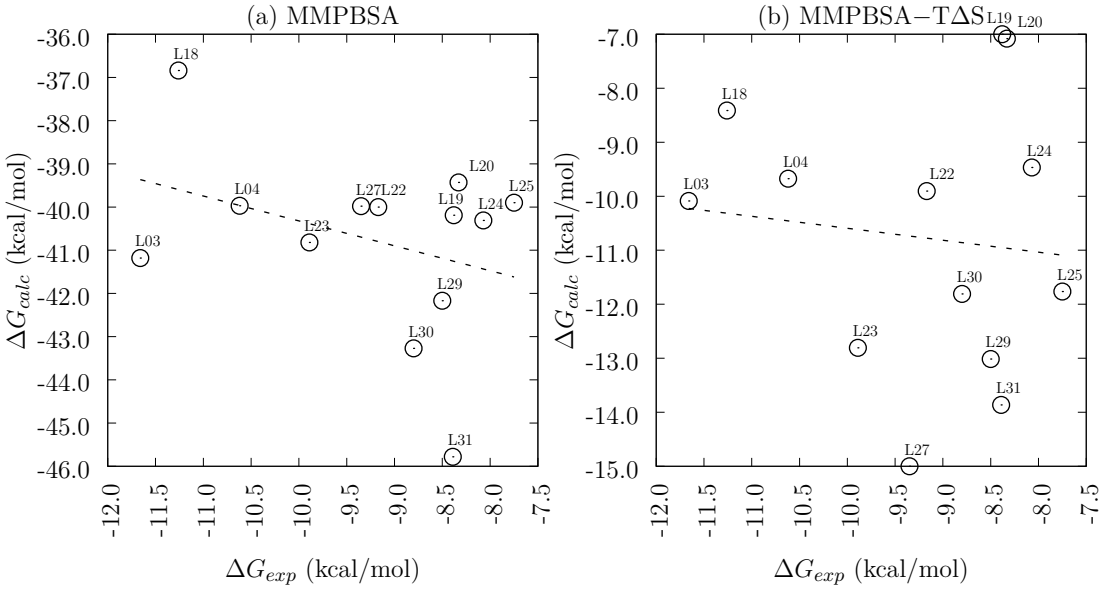


Figure 5.1: *Correlation plot for calculated and experimental ΔG values for 13 ligands complexed to PAK4, using the 1-trajectory ESMACS method: (a) using the MMPBSA free energy method and (b) MMPBSA including the configurational entropy term ($T\Delta S$) generated using normal mode analysis. The dotted line shows the line of best fit. Error bars are removed for clarity, but standard error was no great than ± 1.0 kcal/mol.*

Both simulations states show the formation of a water pocket surrounding the oxygen of ligand L01. A water bridge is formed between this oxygen and a neighbouring serine or lysine residue. The small water pocket in Fig. 5.2a is occupied by three water molecules that are trapped in this location, isolated from the bulk solvent. Conversely, in Fig. 5.2b, where crystal waters were included from the start of the simulation, a water tunnel is maintained which stretches through the binding pocket and is therefore able to interact with the oxygen of L01. The removal of key water molecules results in a change in ligand conformation to accommodate

the absence of the water tunnel. In the simulation with no initial crystal waters, a dihedral rotation is witnessed between the atoms beginning with the first carbon on which the oxygen is substituted, and the first carbon of the benzene ring (C-N-C-C), highlighted in Fig. 5.2c. This rotation causes the benzene and substituted amide to block the small water pocket from being exposed to the bulk solvent. This rotation is not witnessed when crystal waters are included in the simulation, and so this ligand conformation allows for a deep water cavity to be formed.

Here we see that the exclusion of crystal water molecules results in critically incorrect changes to the binding pocket and ligand conformation. Crystal waters are required to maintain the structure of the water tunnel, which keeps the ligand in the conformation favoured in reality. The absence of these waters causes the ligand to adopt a different, and more stable conformation, resulting in a more energetically favourable complex. The inclusion of crystal waters correct this, maintaining the expected ligand conformation. It is essential to understand fully the binding pose of any ligand-receptor complex to be able to execute correct modelling.

Ligand L01 shares little structural relation to the remaining ligands whom all belong to a congeneric series. This also means that L01 adopts a slightly different binding pose. Although both groups of ligands have the same pyrazolyl pyramidine amine moiety, L01 extends from the pyrazolyl end to make key interactions with lysine and asparagine residues. The congeneric series extends from the opposing pyramidine flank, interacting with asparagine and aspartic acid. Due to the structural differences, L01 is redundant for our investigation to elucidate the reason for a lack of correlation within the congeneric series (Fig. 5.1). The remainder of this study will focus on the results obtained from simulations with crystal water molecules.

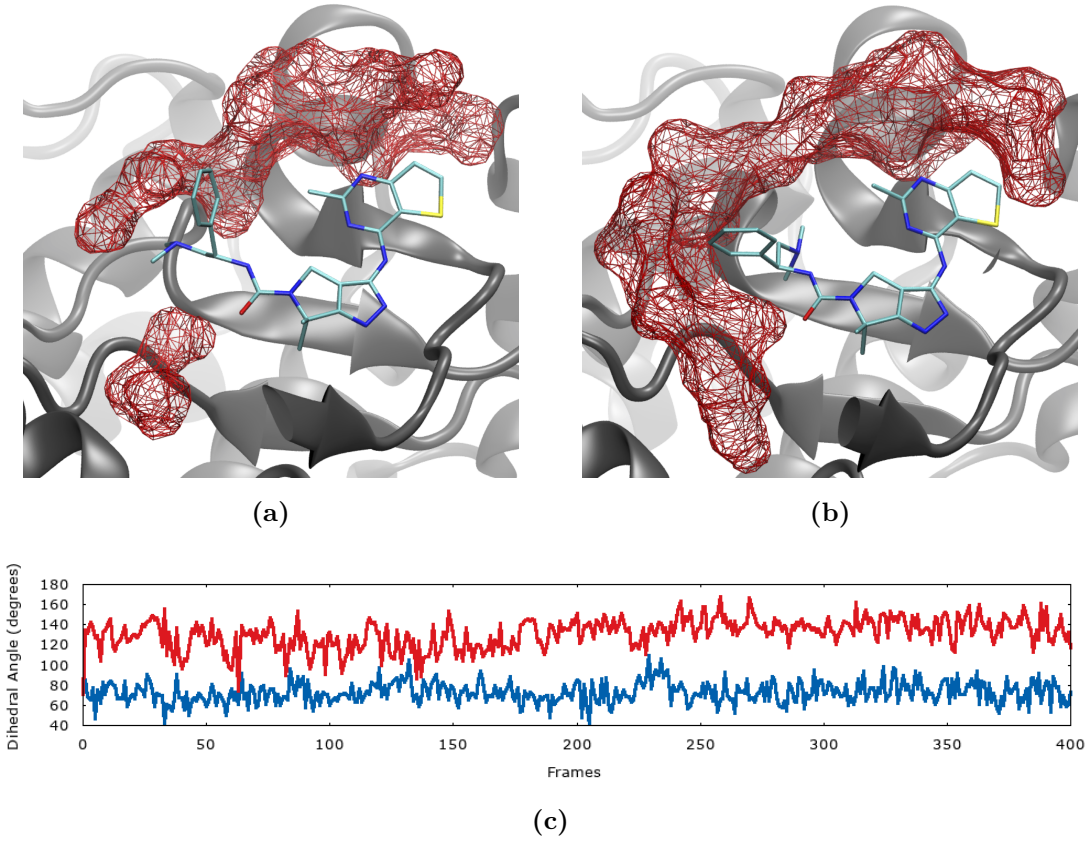
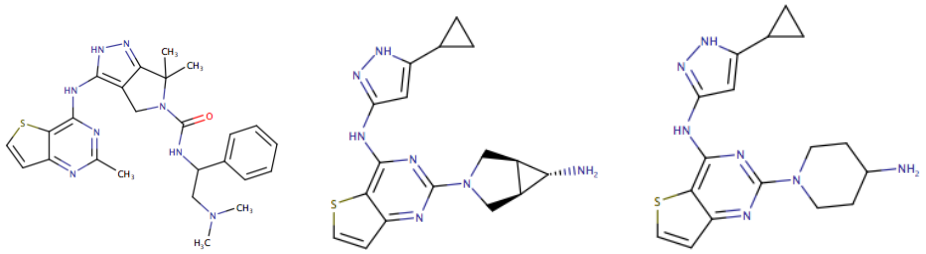


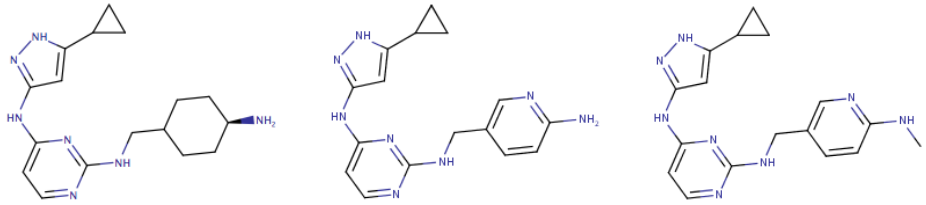
Figure 5.2: *The rotation of a dihedral angle results in the ligand tail regulating the exposure of the water pocket (represented by a red wired frame) to the bulk solvent; (a) simulations that excluded crystal waters show that an isolated water pocket is formed due to a change in ligand conformation, (b) when crystal waters are included at the start of the simulations, the water tunnel that characterises the binding pocket is maintained, and the conformational integrity of the ligand is subsequently maintained, (c) the absence of a water pocket causes a change in ligand conformation through a rotation of a bond; with crystal waters (blue line) this dihedral angle averages ca. 70° and without (red line) it is ca. 140°.*

5.3.2 Aliphatic cyclic moieties result in aberrant electrostatic free energies

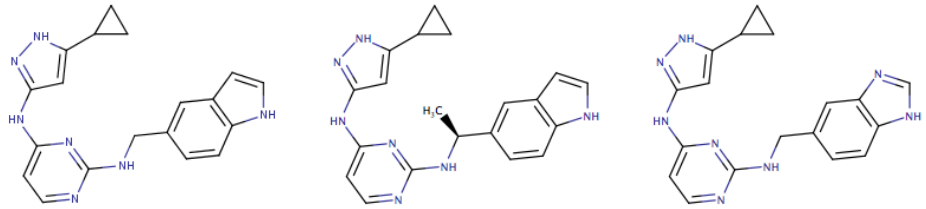
Observing the chemical structures of the congeneric series, we see that the variable group of ligands L03, L04 and L18 possess unsaturated rings, compared with the remaining ligands which all have saturated, aromatic rings. The trends observed



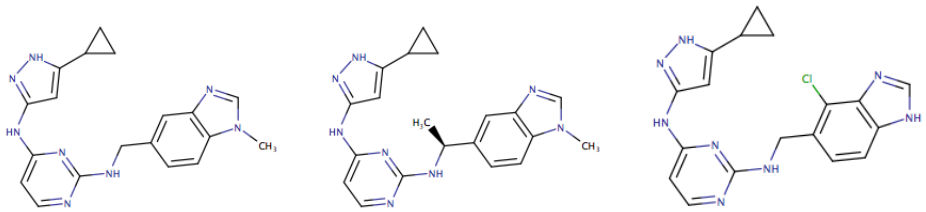
(a) L01 *-10.49 kcal/mol* **(b) L03** *-11.66 kcal/mol* **(c) L04** *-10.62 kcal/mol*



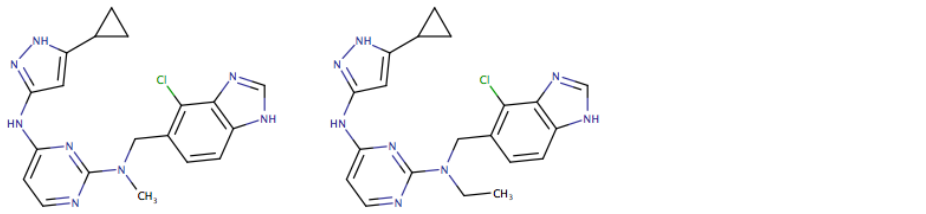
(d) L18 *-11.26 kcal/mol* **(e) L19** *-8.38 kcal/mol* **(f) L20** *-8.33 kcal/mol*



(g) L20 *-8.33 kcal/mol* **(h) L23** *-9.89 kcal/mol* **(i) L24** *-8.07 kcal/mol*



(j) L25 *-7.75 kcal/mol* **(k) L27** *-9.35 kcal/mol* **(l) L29** *-8.50 kcal/mol*



(m) L30 *-8.80 kcal/mol* **(n) L31** *-8.39 kcal/mol*

Figure 5.3: *Chemical structures and binding affinities of 14 PAK4 inhibitors.*

between the two ligand sub-sets were assessed more closely. Decomposed free energy terms were obtained to identify any differences between the two sub-sets and thus allow us to make comparisons between the two sets. Indeed, the existence of a sub-set of ligands was confirmed; the L03 L04 and L18 sub-set corresponds to a less negative value for the electrostatic free energy contribution associated with binding (Fig. 5.4b).

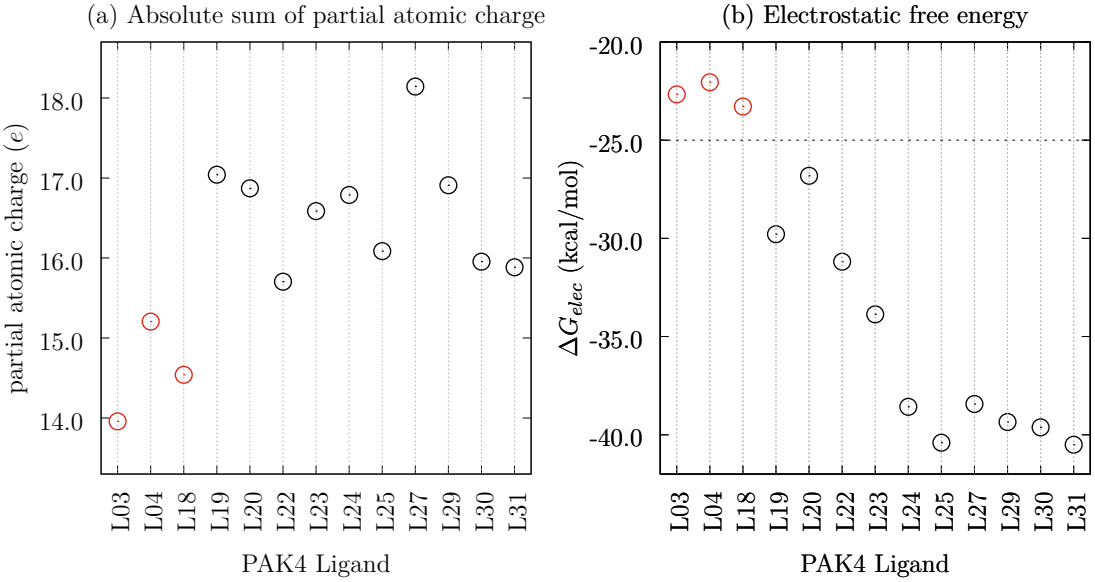


Figure 5.4: *Distribution of: (a) the absolute sum of the partial atomic charge, and (b) the calculated electrostatic free energy associated with the binding free energy, for 13 PAK4 ligands.*

Electrostatic interactions are defined by the non-bonded interactions between partially charged atoms in the ligand and surrounding residues, and so analysis of the partial charges of the ligands were performed. To do this, the sum of the absolute values of the partial charges of each atom of the ligand was calculated. If the sum of the partial charges of a ligand is higher compared to other ligands, this means that the partial charges on the atoms of the variable groups are, relative to the other ligands, more highly charged. This means that ligands with a large absolute sum will consequently contribute more to the electrostatic free energy term of the binding free energy. Conversely, if the absolute sum is relatively low, the electrostatic free energy term will have a smaller contribution. Large fluctuations in these

values would indicate that the electrostatic contribution to binding may be over or under-estimated in the aliphatic sub-set, compared with the aromatic ligands. The total absolute sum of partial atomic charges for the ligand (Fig. 5.4a), shows that the charges are generally weaker, and so do not form comparatively strong electrostatic interactions with surrounding residues.

The knowledge that errors in the electrostatic free energy of binding most likely lead to poor binding affinity predictions, has led to discussion about ways to repair the electrostatic free energy predictions. The first possible problem is that ESMACS is using an implicit water model that does not take into account key solvent interactions between the ligand and the surrounding residues. Inclusion of explicit water molecules at the back of the ligand pocket would potentially result in a more accurate representation of the electrostatic changes in free energy due to binding. Secondly, the internal dielectric constant values are not clearly defined within the MMPBSA approach and so errors in free energies could be as a result of incorrect assignment of internal dielectric constants. Finally, although the aforesaid assessments do not categorically attribute the deviations in electrostatic free energies to poor selection of ligand parameters (and subsequent partial charge assignment), it must be noted that force field selection, and the manner in which partial atomic charges are generated, can play a significant role in final calculated binding free energies. We then proceed to interrogate two ideas: (a) the inclusion of explicit water molecules, and (b) the modification of the internal dielectric parameter.

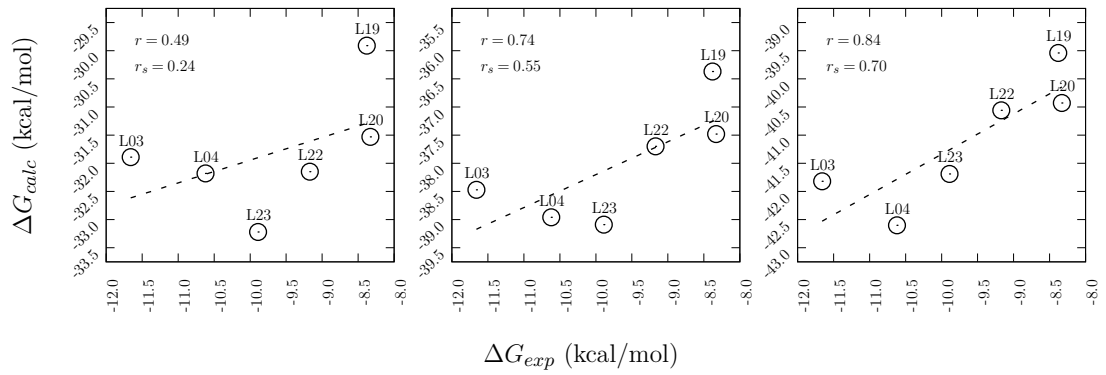
5.3.3 MMPBSA calculations with explicit waters

Binding affinities calculated using the MMPBSA method, with the inclusion of explicit water molecules, are explained in section 5.2.2. This approach resulted in marginally improved correlation. The inclusion of 20 explicit water molecules was the upper limit in performance. Although these results are consistent with the study done by Maffucci et al. [168], the improvements in ranking, in this case, are

not as significant. The incorporation of this method seems limited, and we have seen previously that the inclusion of selective explicit water molecules are most effective when performed manually [16].

5.3.4 MMPBSA calculations with modified internal dielectric

Investigation of the internal dielectric began by calculating binding affinities of 6 ligands using different dielectric constants (Fig. 5.5). Free energy calculations using MMPBSA were only performed on the systems containing crystal waters, with ϵ_{int} values of 1, 2 and 4. The GB form has been omitted. This is because ϵ_{int} in the GB equation is the medium from which the complex is being transferred from (i.e. vacuum), so modifying ϵ_{int} to anything other than 1 would mean the solvation free energy is calculated incorrectly. In the PB form, the internal dielectric is the dielectric of the solute, and so modification of this is theoretically acceptable.



(a) $\epsilon_{int} = 1$ (b) $\epsilon_{int} = 2$ (c) $\epsilon_{int} = 4$

Figure 5.5: *Correlation plot of 6 PAK4 ligands with different ϵ_{int} values using the 1-trajectory ESMACS approach. Only the MMPBSA free energy method, without configurational entropy is reported here. As the ϵ_{int} increases from 1 to 4, the correlation and ranking improve significantly. r_s and r_p are the Spearman rank and Pearson correlation coefficients and ϵ_{int} is the internal dielectric constant. Standard error was no greater than ± 1.0 kcal/mol.*

We see that an increase in ϵ_{int} is followed by a considerable improvement in corre-

lation and ranking. These promising improvements suggest that changing the ϵ_{int} will bring better correlations than using default settings. Studies show that the non-polar environment of the PAK4 binding pocket – no regular polar interactions occur other than the typical 3 hydrogen bond configuration seen in kinases – would benefit from increased ϵ_{int} values [137], and this is supported here. However, there is still a lack of a protocol that can systematically assign the correct ϵ_{int} for a given system, based on the polarity of solute. At least for the PAK4 system, a ϵ_{int} value of 4 gives strong correlations.

From this, we hypothesised that ϵ_{int} is a major factor in predictive performance of binding affinities. To strengthen this claim, the remaining ligands, bar L01, were introduced to the study and binding affinities were computed for $\epsilon_{int} = 1$ and 4, using the MMPBSA approach.

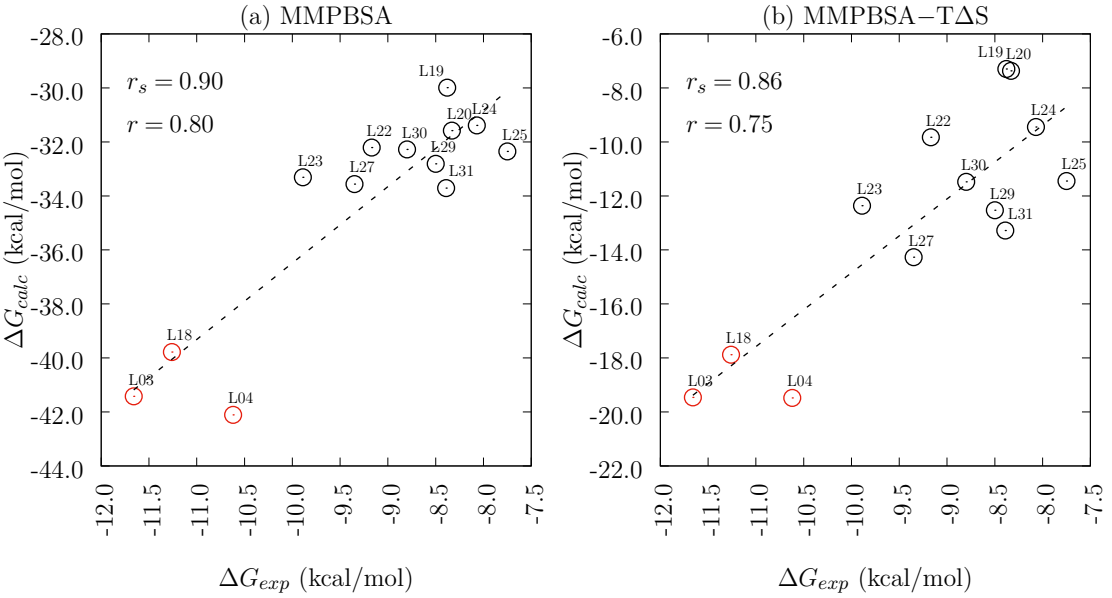


Figure 5.6: *Correlation plot for calculated and experimental ΔG values for 13 ligands complexed to PAK4, using the 1-trajectory ESMACS method: (a) using the MMPBSA free energy method and (b) MMPBSA including the configurational entropy term ($T\Delta S$) estimated by normal mode analysis. The dotted line shows the line of best fit. The red data points are ligands L03, L04 and L18 that are associated with ϵ_{int} value of 4. The black data points have been assigned an ϵ_{int} value of 1. Error bars are removed for clarity, but standard error was no greater than ± 1.0 kcal/mol.*

Conversely to our hypothesis, modification of the ϵ_{int} parameter saw no ranking or correlation with experimental values. However, when ϵ_{int} was changed to 4, for the aliphatic group of ligands, and kept at 1 for the aromatic group of ligands, we report a correlation coefficient above 0.80, and a ranking coefficient above 0.90 for all trajectory approaches (Fig. 5.6). At this stage, the configurational entropy term was included, but we see a slight decline in correlation and ranking. As a result, we suggest that the modification of the ϵ_{int} is ligand-specific, rather than system specific. This stays true to the theory that a non-polar environment would benefit from increased ϵ_{int}, and perhaps the less polar environment associated with the aliphatic sub-set of ligands, requires a different ϵ_{int} compared to the remaining ligands.

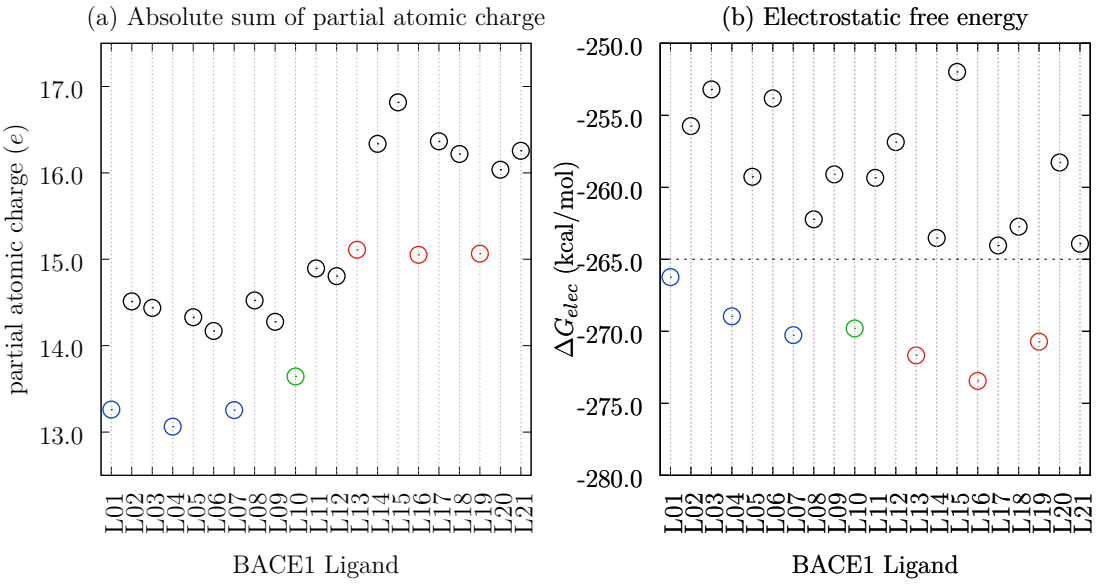


Figure 5.7: *Distribution of: (a) the absolute sum of the partial atomic charge, and (b) the calculated electrostatic free energy associated with the binding free energy, for 21 BACE1 ligands.*

The trends described above were investigated further in the BACE1 system. The BACE1 ligand data set contains ligands with varying ring types and provides a good comparison with PAK4. The data set is comprised of 21 ligands which can be sub-divided into 3 categories: ligands containing 5-, 6- and 7-membered rings. Within each sub-set, the ligands are distinguished through a number of different

substitutions at the variable group (Fig. A.7).

The BACE1 system shows an opposing trend to what is seen for PAK4. The ligands containing a 5-membered ring have, comparatively, a smaller total partial charge (Fig. 5.7a). Based on the PAK4 observations, we would expect these ligands to have the most positive ΔG_{elec} but in fact, relative to the related ligands with 6-, and 7-membered rings, they have the most negative ΔG_{elec} (Fig. 5.7b). Increasing the ϵ_{int} to 4, as we did in the PAK4 study effectively dampens the ΔG^{MM}_{elec} and ΔG^{X}_{sol} terms, from which we obtain a better representation of calculated binding affinities, with respect to experimental results. In the case of BACE1, we amplify the ΔG^{MM}_{elec} and ΔG^{X}_{sol} terms, by modifying ϵ_{int} with varying degrees, to obtain a ΔG_{calc} more representative to what was evaluated experimentally (Fig. 5.8).

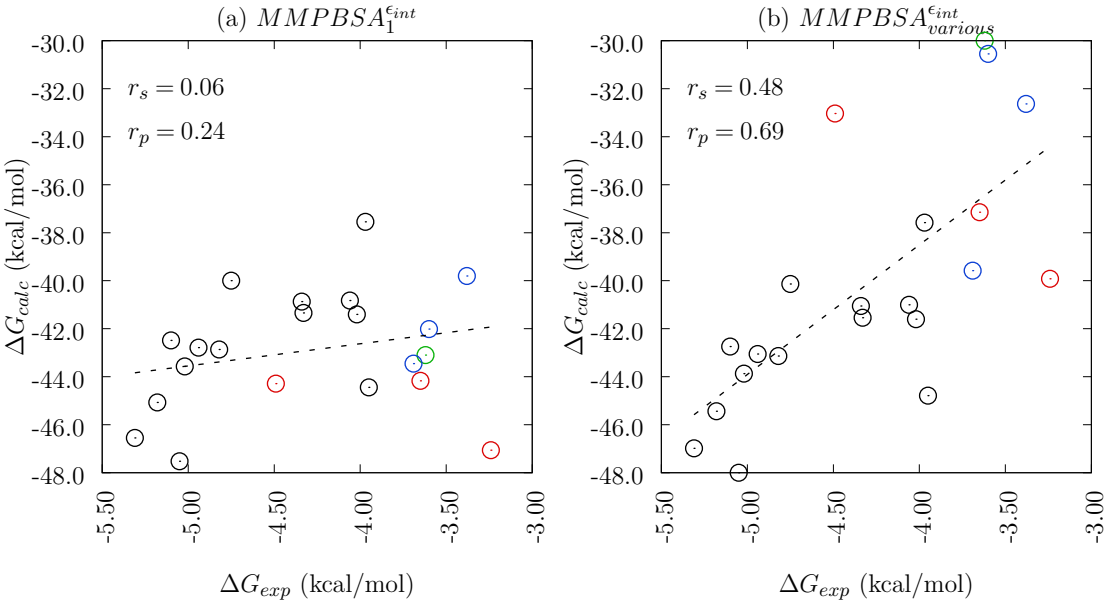


Figure 5.8: *Correlation plot for calculated and experimental ΔG values for 21 ligands complexed to BACE1, using the 1-trajectory ESMACS method: (a) using the MMPBSA free energy method with an ϵ_{int} value of 1, and (b) using the MMPBSA method with varying ϵ_{int} values. The black data points are assigned $\epsilon_{int} = 1$; the blue data points, $\epsilon_{int} = 0.2$; green data points, $\epsilon_{int} = 3$; and red ligands, $\epsilon_{int} = 0.5$. The dotted line shows the line of best fit. Error bars and data labels are removed for clarity.*

It was noticed that modifying the ϵ_{int} parameter according to the relative total

partial charge yields an improved correlation and ranking. For instance, the three ligands with the lowest sum of the total partial charge were assigned an ϵ_{int} value of 0.2 (L01, L04 and L07), whilst the three ligands with the highest sum (L13, L16 and L19) were assigned an ϵ_{int} of 0.5. One ligand that falls in between these two sets (L10), was assigned a value of 0.3.

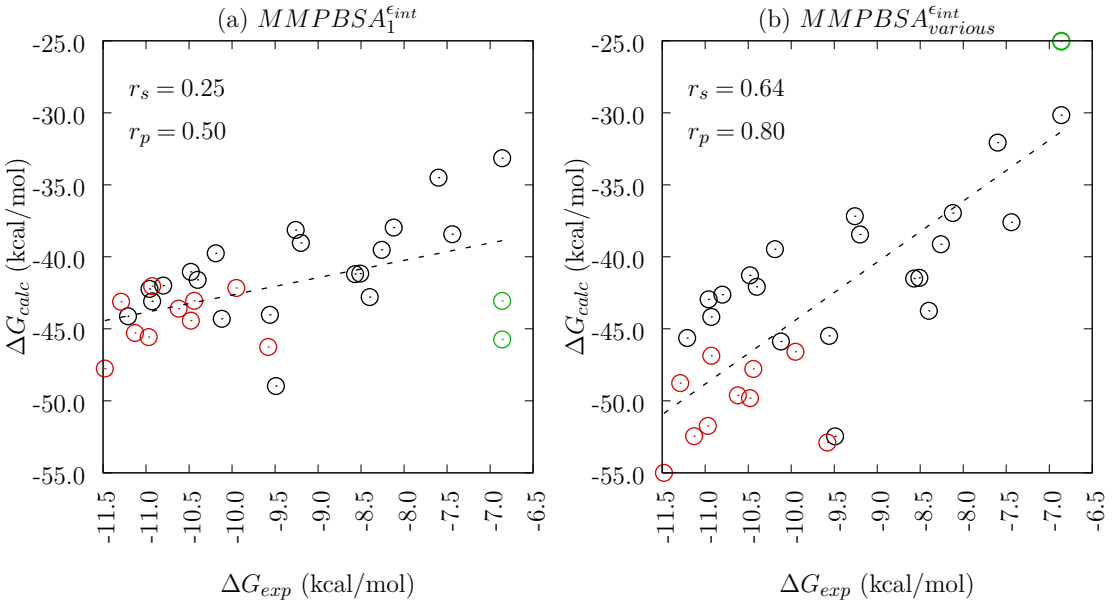


Figure 5.9: *Correlation plot for calculated and experimental ΔG values for 32 ligands complexed to ROS1, using the 1-trajectory ESMACS method: (a) using the MMPBSA free energy method with an ϵ_{int} value of 1, and (b) using the MMPBSA method with varying ϵ_{int} values. The black data points denote $\epsilon_{int} = 1$; the red data points, $\epsilon_{int} = 0.9$, and the green data points are $\epsilon_{int} = 4$. The dotted line shows the line of best fit. Error bars and data labels are removed for clarity*

The obvious reason for this is that in the PAK4 system, the ΔG_{elec} term is over-estimated, and thus requires a dampening effect. Conversely, in the BACE1 system, the said energy terms require an amplifying effect. However, on both occasions, the atomic partial charges are relatively small compared to other ligands in the study, so this, in effect, rules out the possibility that the total atomic partial charges is an important component when deciding the ϵ_{int} value. Thus, in turn, it is unlikely that the over-arching issue is the parameterisation of ligands, and subsequent partial atomic charge assignment, rather, the way in which MMPBSA

calculates the long-range interactions has inconsistencies.

The ROS1 system also sees an improvement in ranking when bespoke ϵ_{int} values are assigned to ligands (Fig. 5.9). After assessing the distribution of the electrostatic free energy of ligands, we assigned a threshold (denoted by a horizontal dotted line in Fig. 5.10). Ligands that posses a more negative ΔG_{elec} value than the threshold, were assigned a ϵ_{int} value of 0.90 to amplify the electrostatic contribution. The remaining ligands, bar L27 and L28, were kept at the default ϵ_{int} parameter.

This selection of ϵ_{int} values yields a r_s value of 0.64, and an r_p value of 0.80 – considerably better than the default ϵ_{int} parameter. The trends observed in the ROS1 system are analogous to the BACE1 system, where an amplification of the ΔG_{elec} results in more accurate binding free energies.

Unlike the PAK4 and BACE1 systems, we see no obvious trends in ROS1 ligands, with regards to the ligand structures to which we can attribute the differences in ΔG_{elec} value. Further, there are no consistencies between the ΔG_{elec} value and the absolute sum of the partial atomic charges (Fig. 5.10), which suggests that there are factors outside the assignment of partial charges, that deliver inconsistent ΔG_{elec} values. Selecting $\epsilon_{int} = 4$ for ligands L27 and L28 generated a binding free energy that aligned well with the regression line, however it is not known why these two ligands require this selection. Similarly to PAK4 and BACE1, ROS1 reports a loss of correlation when $T\Delta S$ is included in the binding free energy.

5.3.5 Assessing a solvent accessible surface area-based entropy method

We proceed to take a closer look at the configurational entropy term in Eq. 2.62. Fig. 5.1 shows that the inclusion of configurational entropy, estimated using NMA (S_{NMA}) results in a less correlated data set. The WSASA method(S_{WSASA}) is introduced here, and compared to the NMA approach. A recent preliminary study

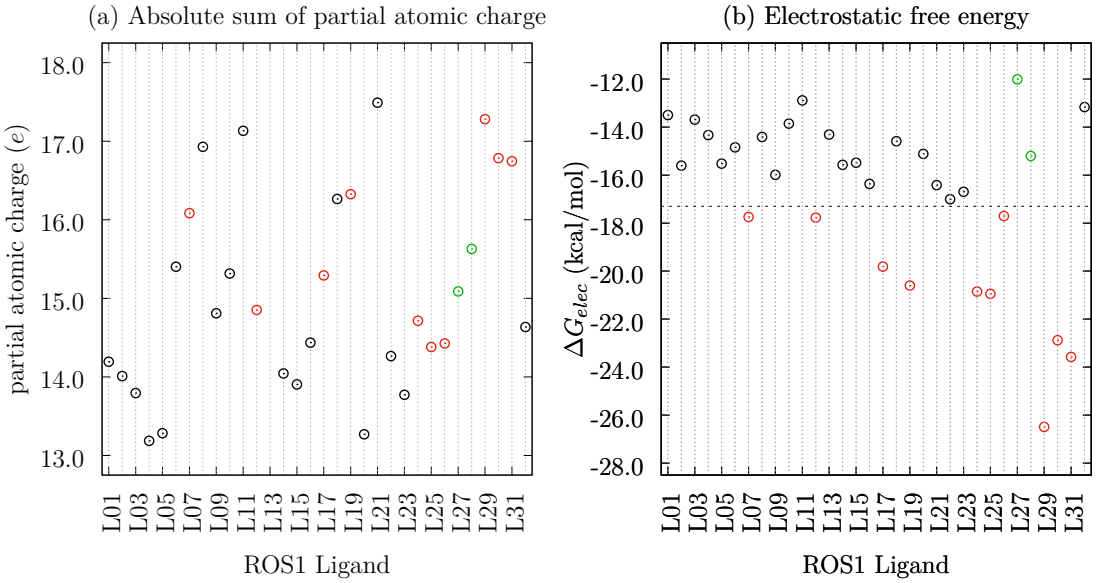


Figure 5.10: *Distribution of: (a) the absolute sum of the partial atomic charge, and (b) the calculated electrostatic free energy associated with the binding free energy, for 32 ROS1 ligands.*

involving WSASA has shown to be able distinguish ligands with different chemical groups. We apply this method in the systems of study in this chapter, to replicate this trend.

A comparison of PAK4 binding affinities in conjunction with the two different entropy estimation methods shows no change in correlation. However, when only ligands containing the aromatic variable moiety are assessed, a significant improvement in ranking and linear correlation are witnessed (Fig. 5.11). Although there is not a global improvement in statistical metrics, we show that WSASA improves the predictive performance when considered with the MMPBSA method. It could be that the calculated binding affinities for this groups of ligands are assigned correctly, and WSASA is correctly estimating entropy values.

The same comparison was performed for the the ROS1 system, and no significant improvement in correlation was witnessed when the WSASA method was incorporated, in fact there was a slight degradation in correlation. The data set was subdivided, and ligands that reported the most negative electrostatic values (red

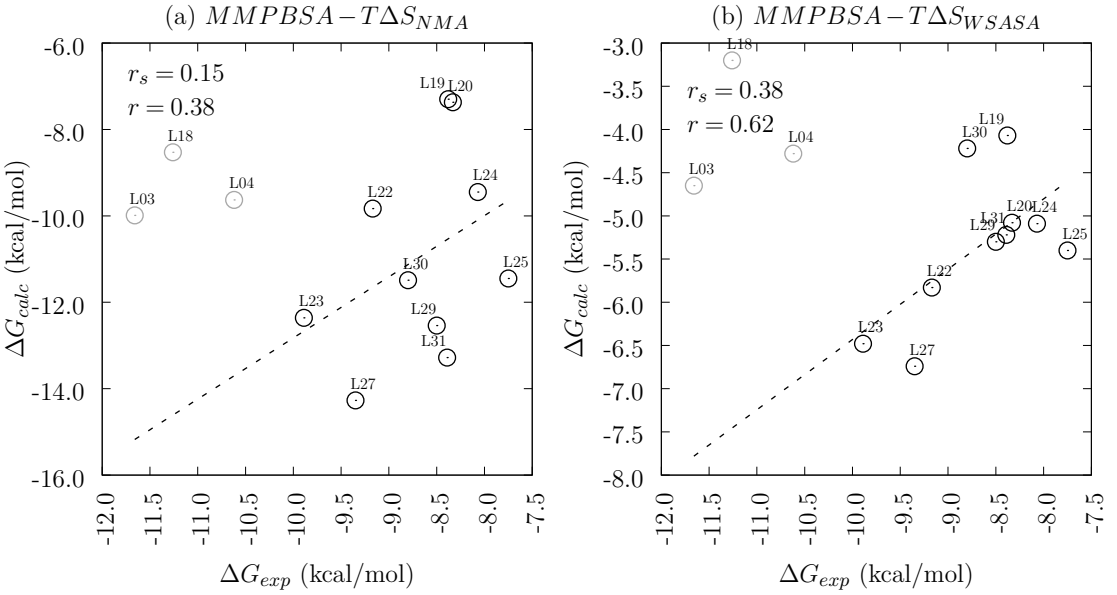


Figure 5.11: *Correlation plots for 13 PAK4 ligands using the 1-trajectory ESMACS approach, and including configurational entropy estimates using: (a) the NMA method and (b) the WSASA method. Data points in grey are L03, L04 and L18 are included to gain a better understanding of the performance of the correlation. Black data points, and consequent regression line are from the remaining ligands in the series. Error bars are removed for clarity.*

data points in Fig. 5.10) were assessed. Here we see a slight improvement in correlation, but not considerable to claim that WSASA improves for the correlation for that particular subset.

With regard to the BACE1 system, it was not possible to obtained configurational entropy estimates using NMA. This is because due to the large system size, an excessive amount of minimising iterations were required to give converged results. Convergence was not achieved for a majority of the analysed frames within the wall clock time of 48 hours. This is indicative of the problem with entropy calculations using NMA. If a system is too large, or far from an energy minima, convergence may not be achievable for some trajectory frames. We were able to obtain configurational entropy using the WSASA method and this reported a degradation in correlation, as seen in the ROS1 example.

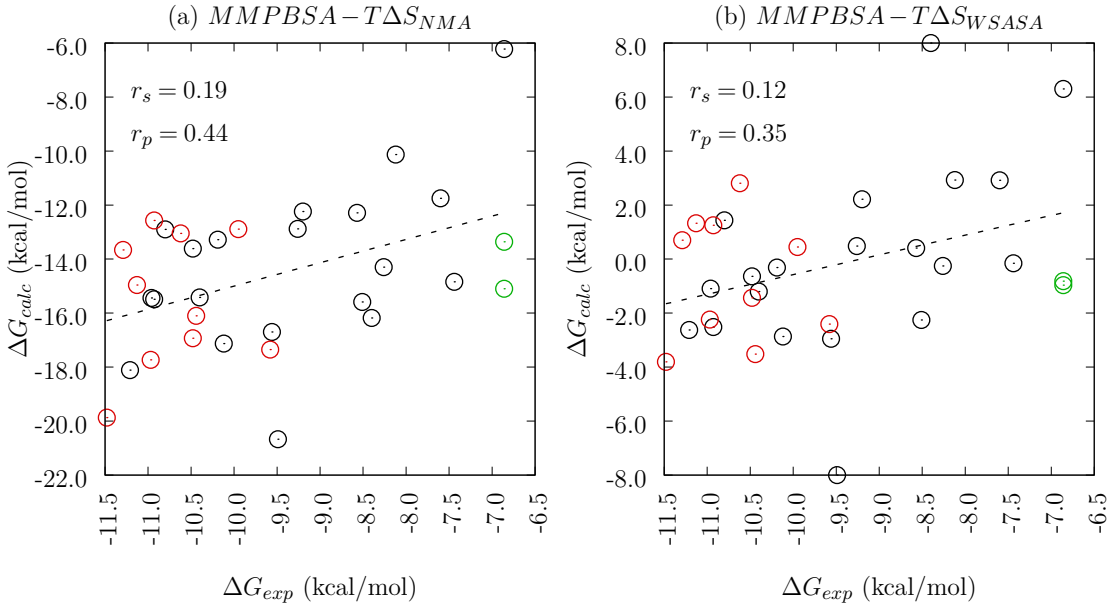


Figure 5.12: *Correlation plots for 32 ROS1 ligands using the 1-trajectory ESMACS approach, and including configurational entropy estimates using: (a) the NMA method and (b) the WSASA method. Data point are coloured respective to the ϵ_{int} selections described previously, however ΔG values used in this plot are obtained using the default ϵ_{int} value. Error bars and data labels are removed for clarity.*

5.4 Discussion

We have shown that, in the PAK4 case study, changing the ϵ_{int} value to 4 for a sub-set of ligands which are relatively less polar, and maintaining default ϵ_{int} values for the remaining ligands, result in vastly improved ranking and correlation in the PAK4 system. This supports previous studies [136, 137, 138] that see large changes in predictive performance when ϵ_{int} is modified. The non-polar residues of the PAK4 binding pocket best describes the improvement in correlation metrics when ϵ_{int} is increased. Maintaining the ϵ_{int} across all ligands results in a lack of correlation or ranking.

A similar trend was seen in the case of BACE1. The most negative electrostatic terms correspond to a subset of ligands which are chemically related. Here, amplifying the electrostatic energy was required. As a result, decreasing the ϵ_{int} in the

range of 0.2 and 0.5 for those ligands, saw a vast improvement in ranking. ROS1 behaves analogously to BACE1 where ligands that fall below a ΔG_{elec} threshold are subjected to an amplification in electrostatic charge, to obtain better rankings. Therefore, we suggest that modification of ϵ_{int} could ligand-specific, rather than system-specific.

Inclusion of explicit water molecules to MMPBSA and MMGBSA calculations result in a slight improvement in ranking, but these are not significant enough for implementation into the ESMACS protocol. Inclusion of more than 20 explicit water molecules, results in no further improvement in ranking coefficients. The purpose of including this approach is that it is easy to implement, and would generate some improvement in rankings. It is indeed easy to implement, however the small improvement in ranking metrics does not qualify this as a method that should be incorporated into all systems of study, where explicit waters are important. Inclusion of explicit waters should be done after gaining a complete understanding of the binding mode of the ligand-receptor complex, and then selecting specific water molecules, as seen in other studies [100]. With this said, there is always a trade-off and the amount gained by including explicit waters is more than off-set by the very system-specific requirements that render the approach must less automatable.

When combining the inclusion of explicit water molecules and a variable dielectric constant we see no improvement in correlation and ranking. It is therefore not suggested to perform the two methods in tandem. Attempting to adjust the ϵ_{int} parameter initially, will be easier and, in this case, generates better improvements in results. Ultimately, valid departures from default settings need to be qualified and at this moment, there is no definitive criterion that can be implemented. On the other hand, one may study the partial atomic charges (similarly to what is seen in Fig. 5.4, 5.7 and 5.10) and distinguish the ligands qualitatively. Further, obtaining binding free energies by changing the MMPBSA ϵ_{int} parameter is negligible in time and effort. In fact, a good estimate can be obtained by dividing the electrostatic

contribution to the binding free energy (ΔG_{elec}^{MM}) and the free energy of solvation (ΔG_{sol}^{X}) by the ϵ_{int} value.

We also highlight the importance of including crystal water molecules prior to simulation. Ligand L01 demonstrates large changes in conformation, and subsequently calculated binding affinity, when crystal waters are removed. This observation is well known, and widespread resulting in several software applications [169, 170] that automate the inclusion of important water molecules.

Using the WSASA method to estimate configurational entropies led to a considerable improvement in ranking in the PAK4 system containing the aromatic series of ligands. No change in ranking was observed for the aliphatic series. WSASA performed equally to NMA when all ligands were taken into account. WSASA was inferior to NMA in the ROS1 data set. Although WSASA has not reported consistently improved results, the ability to improve rankings for aromatic ligands in PAK4 is noteworthy. The ease of completing WSASA calculations on a desktop computer at a fraction of a time, compared to NMA, means that it is a viable tool for quick entropy calculations based on actual trajectory frames. The BACE1 system demonstrated the difficulties associated with NMA, and WSASA circumvents this issue.

To conclude, we have shown that varying ϵ_{int} values make a significant difference in the quality binding affinity predictions when the ESMACS protocol is employed. Although a quantitative 'rule' for ϵ_{int} selection has not been formally outlined in this chapter, it is evident that appropriate treatment of this term is critical to achieve correct binding affinity values. The selection of ϵ_{int} will most certainly depend on the polarity of the binding pocket of the recepter and ligand in question, and so cannot be generically defined.

Chapter 6

Application of ensemble-based binding affinity protocols in a clinical setting

6.1 Introduction

Much of this thesis so far has described the need for rapid and reliable binding affinity predictions in a drug discovery setting, however, ensemble-based binding affinity protocols can also be applied in the medical domain. Today, genotypic assaying (or genome sequencing) is common practice. Genome sequencing technology uncovers the genetic information of a patient or cohort of patients that are suffering from a genetic disease. From this genetic profile, clinicians can deduce the genetic nature of the disease and prescribe the appropriate therapy [171]. Although significant advances have been made in genome sequencing capabilities, obtaining a genetic profile of a patient and subsequently choosing the correct diagnosis, is a non-trivial approach meaning that clinicians have often turned to decision-support tools. These tools are aimed for the early diagnosis of patients, allowing the correct therapy selection. Unfortunately, decision-support tools have also been found

wanting, and there is a demand for an accurate and reliable tool that can efficiently diagnose patients. Such methods must gain regulatory approval and so these must be reliable methods. With this said, there is currently no truly clinical applications for disease prediction, or early diagnosis. In its infancy, the ESMACS approach has been successfully employed to determine binding affinity of HIV-1 protease inhibitors that are currently used in the clinic, based on different protein sequences [43, 87].

Half a century after the discovery of the double helical structure of DNA, the completion of the Human Genome Project (HGP) paved the way for a new era in genomic research. Relatively crude, first-generation methods were used to sequence the human genome during the early stages of the HGP, of which Sanger sequencing [172] was the most prominent. Following the inception of Sanger sequencing, a number of improvements were made in the following years. Flourometric-based detection and capillary electrophoresis allowed for reactions to occur in a single vessel (previously performed in 4 vessels), and improved the capability to detect nucleotides more accurately. These advances resulted in the development of commercial DNA sequencing machines [173], and eventually simulataneous sequencing machines thare are able to process hundreds of samples [174, 175].

The rate at which nucleotide sequencing technologies are developing resulted in a 'genomic revolution' that is growing at a faster rate than the 'computing revolution' [171]. This can be compared using Moore's law, which states that the complexity of microchips (measured by the number of transistors per unit cost) doubles every 18 months. The capabilities of genome sequencers has grown at a faster rate than Moore's law, doubling every five months between 2004 and 2010. Today next generation sequencing (NGS) methods [176, 177, 178] give researchers a plethora of diverse tools from which sequencing can be conducted efficiently. Arguably the most popular approach is the Solexa method of sequencing developed by Illumina [179]. The implications of this are faster and cheaper methods for gene sequencing.

In 2006, Illumina's first sequencer could sequence a genome for $300,000. A decade later, this has dropped to $1,000 [180]. Illumina claim that their new machines are 70% faster and within 3 to 10 years, they will be able to sequence a genome for $100. Oxford Nanopore is an emerging method that can read longer lengths of genetic code, making it very efficient [180]. The cost of the Nanopore machine is considerably cheaper ($75,000) than some of Illumina's products, which can cost close to $1,000,000.

NGS techniques are not limited to the mutations in the human genome; nucleotide sequencing can also be applied to pathogens – any infectious agent that can cause disease such as a virus, bacteria, fungi [181] etc. – which has a genetic basis. In 2015 the Ebola virus epidemic was a result of a mutation in the genetic code of *Ebolavirus Zaire evolavirus*. Within two days of collecting the sample, researchers were able to sequence the viral genome [182]. As such, genomic research in terms of pathogenesis can be exploited in two ways: understanding the human genetic variation related to health and disease, and uncovering the genetic profile of the pathogen itself.

We are now in an era where we are able to obtain complete DNA sequences of various cancer genomes, and with that comes the discovery of ever more mutations that result in oncogenesis, and resistance to cancer therapies [183, 184]. Our growing understanding of the cancer genome also demands reliable and rapid decision-support software to determine the correct therapy based on a patient-specific genetic profile. Thus, an imperative goal in the medical domain is the development of a reliable and high-throughput computational tool that is able to rank the potency of current therapies across a range of mutated receptors, on clinically relevant timescales. The "INtegrated and Scalable PredictIon of REsistance" (INSPIRE) project, a US Department of Energy (DOE) funded 'INCITE' award addresses the challenges represented by drug resistance, through the use of large scale atomistic molecular simulation [185]. This initiative aims to guide precision cancer therapy in which

anti-cancer treatments are tailored specifically to patients based on the genome of their particular cancer.

The work described in this chapter is a collaborative effort between researchers at University College London (UCL), Carnegie Mellon University in Qatar (CMU-Q) and Hamad Medical Corporation (HMC), funded by the Qatar National Research Fund (QNRF). In the state of Qatar breast cancer consitutes 39% of all cancers in females [186]. Consequently it is pertinent to research how breast cancer develops in this population to design an appropriate method of treatment. To develop such a treatment, it is critical to have a comprehensive understanding of the genetic mutations that cause breast cancer. Gene sequences of the estrogen receptor (ER) will be completed by scientists from HMC and CMU-Q, and will be used in molecular modelling studies at UCL. Specific mutations, if any, in the Qatari populations will be investigated for their effects on the drug binding affinities.

We begin by aiming to understand how the ensemble-based protocols employed throughout this thesis perform for the wildtype (WT) ER and 2 mutated variants, Y537S and D538G, which are the most common in breast cancer patients [187]. Absolute binding affinities, using ESMACS, will be computed for 5 current cancer therapies, and the hormone estradiol which is an endogenous agonist of the ER. ESMACS binding affinities will be generated for the 6 ligands bound to the three respective ER variants. The TIES protocol is implemented for 3 transformations, of which all are inhibitors of ER. Similarly to ESMACS, TIES relative binding affinities will be generated for the WT and ER mutants.

6.2 Hypothesis

In the following sections we will see that the ER undergoes large conformational changes, dependent on ligand binding and the presence of mutations on the receptor. Hence, the decision to investigate ER is on the basis of the receptor flexibility. The conformational variation seen in the ER can be difficult to model, particularly

in short simulations that are implemented in ESMACS and TIES. The ESMACS method comes with different settings, namely 1-, 2- and 3-trajectory ESMACS approaches, described in section 2.6.2.3. The 3-trajectory approach calculates free energies of the three entities (complex, receptor and ligand) from individual simulations and then takes the difference (Eqn. 2.64c) to achieve the binding affinity of the ligand to the receptor. The benefit of this approach is that in the free simulations of each entity, additional conformations are sampled compared to the 1-trajectory approach, where the ligand and protein are both restricted in movement due to being in the bound state. The 3-trajectory ESMACS approach, then, would be suited to receptors that undergo large conformational changes upon ligand binding. We therefore hypothesise that the 3-trajectory ESMACS approach will provide us with a better estimate of the binding free energy of ER and its mutant variants, using the six ligands shown in Fig. 6.3, due to the conformational flexibility of the ER receptor upon ligand binding. We will use TIES relative binding affinities to support this hypothesis.

6.2.1 Role of the ER in breast cancer

ERs are ligand-activated nuclear transcription factors which bind estrogens to induce physiological effects [188]. Estrogens are hormones that bind to the (ER) and induce transcriptional regulation, which could either be activation of a particular target gene or repression, in conjunction with interacting transcription machinery [189, 190]. The ER regulates the differentiation and maintenance of various different tissues [191, 192]. In breast cancers, ER is a major contributor to cell growth, survival and metastasis [187]. There are two types of ER, ER-α and ER-β, however ER-α plays the prominent role in ER-positive breast cancer.

ER ligands (Fig. 6.3) bind to the ligand binding domain (LBD) located on the C-terminal. Due to the flexibility and spaciousness of the cavity, a diverse range of small molecules are able to bind in the LBD, triggering many different physiolog-

ical effects. The endogenous estrogen hormone, estradiol, interacts as an agonist (binds to ER to induce a biological response), whereas synthetic compounds such as tamoxifen exhibit antagonistic behaviour (binds to ER and inhibits a biological response). The latter belongs to a class of drugs called "selective estrogen receptor modulators" (SERMs) which have been used extensively to treat estrogen receptor-positive breast cancer. Estradiol is sometimes written as 17β-estradiol. Here, we will simply use estradiol for convenience and clarity.

6.2.2 Effects of mutations on ER function and drug efficacy

Targeted hormone therapy represents a major advance in the treatment of human cancers which specifically inhibit selected molecular targets. About 2 out of 3 breast cancers are hormone receptor-positive (that is estrogen- or progesterone-receptor positive), in which high hormone levels induce carcinogenesis. Hormone therapy can reduce the onset of cancer by either lowering hormone levels or blocking hormones from binding to their receptor targets and thus preventing a biological response. Tamoxifen is such a drug that blocks ERs in breast cancer cells and has been used clinically for more than 30 years to treat hormone receptor-positive breast cancer.

Although tamoxifen has been used successfully for 3 decades, ER drug resistance to tamoxifen has resulted in a less efficacious therapy. Drug resistance comes in two forms: *de novo* resistance and acquired resistance. The former is found in ER-positive patients who are unresponsive to hormone therapy from the beginning of treatment, while the latter is developed in those ER-positive patients who are initially responsive to hormonal treatment but then acquire resistance. *De novo* resistance, although rare, is a result of a genetic disposition of the patient to mutations in the ER [193]. Considering the clinical value of tamoxifen in treatment of breast cancer, it is clear that preventing and overcoming tamoxifen resistance remains an important clinical goal.

At this point it is essential to mention that tamoxifen is not an active, therapeutic compound, rather, once tamoxifen is passed through the liver, it is metabolise into afimoxifene which is responsible for the therapeutic effect. Tamoxifen is metabolised into afimoxifene, which is the main active metabolite and endoxifen, a minor metabolite. This distinction is important because in this chapter binding affinities calculations are performed for all three molecules.

6.2.3 Agonist and antagonist binding to ER

Binding of ligands to the ER is facilitated by heat shock protein 90 (HSP90, [194]) which, in the absence of a ligand, opens up the steroid binding cleft to allow entry of a ligand [195]. HSP90 act as a chaperone, which also expedites the formation of ER dimers (two ERs covalently bound together), and recruitment of coactivators, to produce an active complex primed for a physiological response. The agonist and antagonist binding [188] within the LBD is described below and visually presented in Fig 6.1.

Estradiol contains a polar moiety forming a small polar sub-pocket within the LBD, which is defined by the phenol group forming hydrogen bonds with glycine (G353), arginine (R394) and a water molecule. This is additionally stabilised by another hydrogen bond with histidine (H524) and an oxygen, on the opposite end of the ligand. The non-polar nature of the main body lends itself to submerging into the pocket and is surrounded by hydrophobic residues. Access to the pocket is restricted by alanine (A350), leucine (L387) and phenylalanine (F404), forming a rigid architecture. Consequently, the structural pre-requisite for ligand entry into the binding cavity is a planar ring which is able to slide past the 'gatekeeper' residues.

When the LBD is occupied by the agonist estradiol, the C-terminal chain, also known as the H12 helix, sits over the ligand binding cavity and forms a 'lid', without directly interacting with estradiol. The charged H12 helix is positioned as

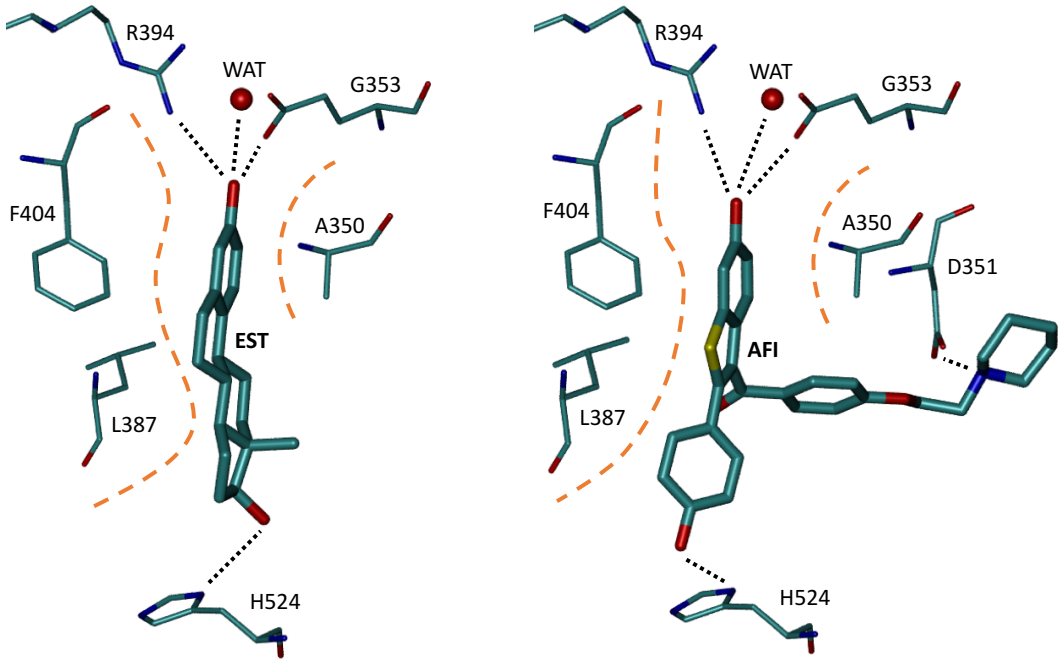


Figure 6.1: *A representation of agonist and antagonist binding in a wildtype ER receptor, using the example of estrogen (left) and afimoxifene (right), respectively. Only key residues that play a role in ligand binding are displayed, and all hydrogen atoms have been removed for better clarity. Hydrogen bonds are shown as black dotted lines. The dashed orange lines indicate residues (A350, L387 and F404) that act like 'gatekeepers' by restricting access to the LBD. 'WAT' is a water molecule.*

such that the highly conserved hydrophobic residues face inwards, toward the LBD, and the hydrophilic residues interact with the solvent. The position of H12, when bound to an agonist, is a pre-requisite for transcriptional activation by generating a AF-2 cleft that is required for interaction with steroid receptor coactivator 1 (SRC-1, [188, 196]) and (SRC-3, [197]).

In the case of antagonists, such as afimoxifene, the binding mechanism is similar. Hydrogen bonds are formed between the alcohol groups on either end of the ligand. However, the long side chain in afimoxifene causes a non-allosteric displacement of the H12 helix. This is supported by hydrogen bonding between aspartic acid, D351, and the piperidene group located at the terminus of the ligand side chain. As a

result, the H12 helix occupies the AF-2 cleft preventing the SRC-1 coactivator to interact with the ER and trigger transcription activity. Fig. 6.2 summaries the mechanism of agonist and antagonist binding in ER.

6.2.4 The effects of ER mutant receptors on the binding mechanism

The mutations of the tyrosine residue at position 537 to serine (Y537S), and aspartic acid at position 538 to glycine (D538G) are the most common cause of acquired resistance to hormonal therapy in ER-positive breast cancer. The mechanisms of action are described below [187].

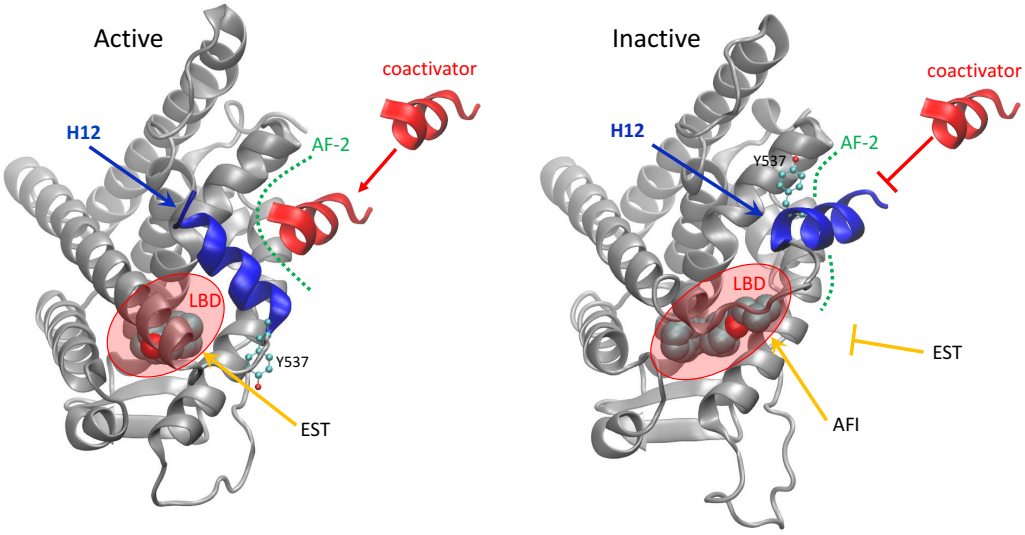


Figure 6.2: *A representation of agonist and antagonist binding in a wildtype ER receptor with estradiol (EST) and afimoxifene (AFI) as examples. During the process of agonist binding, the modular non-polar estradiol settles in the binding pocket allowing the H12 helix to fold over and act as a 'lid' for the LBD. As a result, the ER is available to interact with coactivators SRC-1 and SRC-3 at the AF-2 cleft and induce transcriptional activity. The binding of an antagonist such as AFI, however, means that H12 is displaced and obstructs the AF-2 cleft. Coactivator SRC-1 cannot interact with the AF-2 cleft and is thus inactive.*

In the D538G mutant receptor bound to estradiol, the H12 helix is displaced ac-

companied by a conformational change in Y537. The Y537 residue in the (WT) ER-estradiol complex forms a hydrogen bond with N348 forcing the hydrophobic side chains in the H12 helix to bury inside the LBD. In the ER-estradiol complex that exhibits the D538G mutation, the Y537-N348 hydrogen bond is not formed and the hydrophobic side chains face outwards [187].

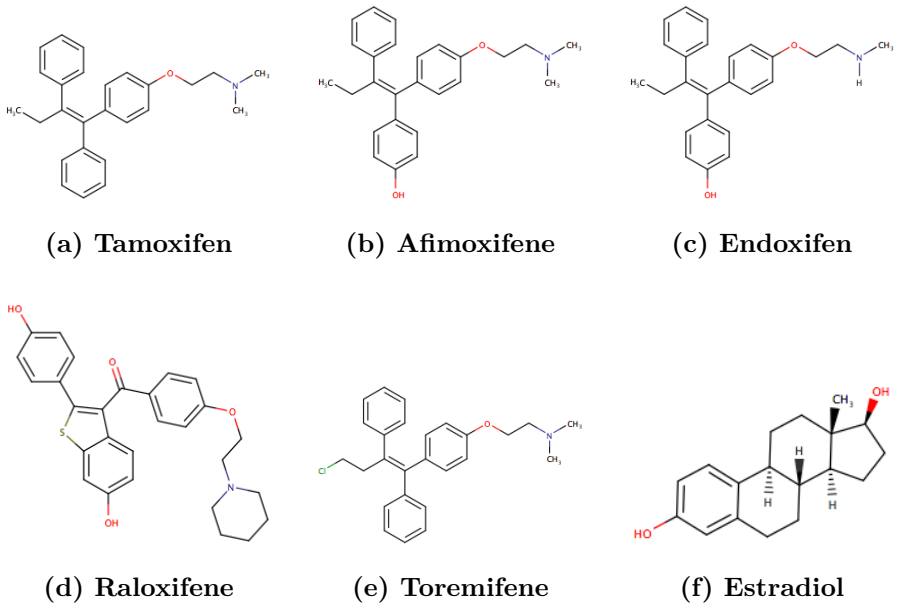
The D538G mutant even shows activity when in the apo (unbound) form. Here, the mutant adopts a similar structure to WT ER-estradiol complex, with the H12 helix twisting inwards towards the LDB. In fact, the H12 helix in the unbound D538G receptor is stabilised further due to the hydrophobic residues burying themselves even deeper into the protein surface [187].

The Y537S also becomes active in its unbound state by forming a hydrogen bond with residue D351. This forms a stable agonist state in the absence of estradiol, and is subsequently ready for interaction with SRC-1 or SRC-3 [188, 196, 187].

6.3 Methods

ESMACS and TIES binding affinity calculations were performed using the standard protocol described in Chapter 2.7. Binding affinities were obtained for 6 ER ligands (Fig. 6.3). Experimental binding affinities were completed using a radiometric ligand-binding assay by Fanning et al. [187, 198] achieving IC_{50} values, which were converted to K_i values using the Cheng-Prusoff equation (2.20). The full and abbreviated names are shown in Table 6.1.

Two ER structures were used for this study, which represent the open and closed forms with the respect to the position of the H12 helix. The PDB code of the open conformation crystal structure is 3ERT [199], and the closed conformation is 1QKU [200]. The ER structure of the former PDB model is complexed to afimoxifene, whereas the latter is bound to the native estradiol. The remaining inhibitors were generated by modifying the AFI inhibitor in 3ERT.



(a) Tamoxifen **(b) Afimoxifene** **(c) Endoxifen**

(d) Raloxifene **(e) Toremifene** **(f) Estradiol**

Figure 6.3: *Chemical structures of 6 ER ligands. Experimental binding affinities are not available.*

Table 6.1: *The full name of the 6 ligands that have been investigated, along with the associated abbreviations employed in this chapter.*

Abbreviation	Full Ligand Name
EST	Estrogen
TOR	Torimefene
EDO	Endoxifen [1]
RAL	Raloxifen
AFI	Afimoxifene [2]
TAM	Tamoxifen

[1] Endoxifen is a minor metabolite of tamoxifen. That is, after tamoxifen passes through the liver it is a small percentage is transformed into endoxifen which imposes a therapeutic effect, not tamoxifen.

[2] Afimoxifene is the major metabolite of tamoxifen, and primarily responsible for its therapeutic effect.

The same crystal structures were used to complete regular TIES calculations. Of the 6 ligands in the study, only 4 qualified for TIES calculations, as they were structurally related. Note from section 2.6.1 that TIES calculations involve an alchemical mutation between two ligands. Excessively large structural differences and/or changes in the charge between two ligands are likely to cause large conformational changes as the ligand mutates, and thus TIES is likely to generate inaccurate relative binding affinities. There is no definitive process for selecting the right TIES pair. Chemical changes to a chemical group within a ligand pair is usually acceptable, for example the removal of a hydrogen atom and the subsequent inclusion of a methyl group is tolerated, like that between endoxifen and afimoxifene. Other considerations are force field parameterisations, with respect to charge variation in TIES ligands. We have seen in other studies pertaining to the CDK2 and TYK2 system that certain chemical groups or structures are not well tolerated in this respect (sulphonamide group and ring structures, respectively). Raloxifene and, in particular estradiol, are structurally too different for this approach, although there is no charge difference between any of the ligands.

The TIES transformations are defined as follows: T01 is an alchemical change from TAM to AFI, T02 is a transformation from AFI to EDO and lastly, T03 mutates from TAM to TOR. TIES relative binding affinities were completed for these transformations bound to the each ER receptor, totaling 9 TIES calculations.

To be able to assess the performance of the binding affinity predictions, it is useful to have experimental binding affinities. Moreover, to perform an accurate assessment against experimental binding affinities, the values need to be obtained using the same method, and in replicates to obtain error bars. There is no account in the literature of experimental binding affinities for the full set of ligands. However, Fanning et al. [187] report experimental binding affinities for EST and AFI bound to WT, Y537S and D538G ERs, respectively. We will use these values to perform a partial assessment of computationally derived binding free energies. Caution must

be taken however, as the reported error bars suggest large uncertainty – results may deviate up to 50% of the binding affinity value with respect to the K_i metric.

ESMACS binding affinities calculations were all performed on the LRZ SuperMUC Phase 1 and Phase 2 machines [105] during a project which utilised the entire HPC facility (approximately 250,000 cores) for a duration of 36 hours (see section 3.4.1). The TIES calculations were performed under normal operational procedures on SuperMUC, and required a total of 610,000 core hours.

6.4 Results

We first assess the binding affinities that have been obtained from ESMACS – namely 1-, 2- and 3-trajectory approaches – against the experimental values reported in Fanning et al. [187]. The results are reported in Table 6.2.

In this analysis, the 3-trajectory ESMACS results show the strongest correlation with experimental values, particularly the PB and PB_{NM} free energy methods ($r_p = 0.92$, $r_s = 0.96$). Conversely, the 1-trajectory approach reports a negative correlation for all free energy methods, since the conformational changes are not sampled in the single trajectory of the complex. Additionally, the inclusion of configurational entropy does not degrade the ranking and correlation like it has done in systems described in previous chapters. The correlation plots of the best performing ESMACS approaches are shown in Fig. 6.4.

The 3-trajectory approach generates binding affinities from separate, but ensemble-based, trajectories of the complex, receptor and ligand. We have learned that upon ligand binding, the ER undergoes large conformational changes, thus, the 3-trajectory ESMACS approach should theoretically capture this resultant free energy change in the receptor. The binding mode of ER ligands to ER, is described by the induced-fit model, which the 3-trajectory approach is able to detect. The 1-trajectory method, for instance, does not take into account the conformational

Table 6.2: *Computational and experimental (ΔG_{exp}) binding affinities of estradiol, EST, and afimoxifene, AFI, to three ER receptors: WT, Y537S and D538G. The computational binding affinities are achieved using the 1-, 2- and 3-trajectory ESMACS approaches. The Generalised Born (GB), and Poisson-Boltzmann (PB) free energy methods were used, in addition to configuration entropy obtained from normal mode analysis (NMA). The Spearman (r_s) and Pearson (r_p) correlations are presented for all methods. All values are in kcal/mol.*

1-trajectory

Ligand	Receptor	GB	PB	GB_{NM}	PB_{NM}	ΔG_{exp}
EST	WT	-28.51	-26.37	-6.86	-4.72	-12.85
EST	Y537S	-28.37	-27.06	-6.64	-5.33	-11.86
AFI	WT	-43.27	-34.58	-20.32	-11.63	-12.70
AFI	Y537S	-42.64	-34.27	-19.99	-11.62	-11.51
AFI	D538G	-42.63	-33.81	-20.16	-11.34	-11.35
Spearman, r_s		0.14	0.15	0.15	0.17	
Pearson, r_p		-0.37	-0.39	-0.39	-0.42	

2-trajectory

Ligand	Receptor	GB	PB	GB_{NM}	PB_{NM}	ΔG_{exp}
EST	WT	-57.82	-56.91	-17.29	-16.38	-12.85
EST	Y537S	-58.50	-57.59	-17.46	-16.55	-11.86
AFI	WT	-38.21	-31.17	-17.51	-10.47	-12.70
AFI	Y537S	-34.04	-26.87	-14.87	-7.70	-11.51
AFI	D538G	-31.73	-23.58	-11.98	-3.83	-11.35
Spearman, r_s		0.27	0.25	0.61	0.41	
Pearson, r_p		0.52	0.50	0.78	0.64	

3-trajectory

Ligand	Receptor	GB	PB	GB_{NM}	PB_{NM}	ΔG_{exp}
EST	WT	-29.93	-40.65	-8.87	-20.23	-12.85
EST	Y537S	-24.16	-30.20	-3.24	-9.29	-11.86
AFI	WT	-43.95	-35.39	-22.07	-13.51	-12.70
AFI	Y537S	-28.86	-18.88	-11.06	-1.08	-11.51
AFI	D538G	-27.84	-16.49	-7.84	3.50	-11.35
Spearman, r_s		0.35	0.92	0.23	0.91	
Pearson, r_p		0.59	0.96	0.48	0.95	

change of the receptor upon binding. All the energy terms cancel out and what is left is the free energy of the direct ligand contacts with the protein. For this reason 1-trajectory ESMACS generates worse rankings and correlations for this ligand set. The promising correlations observed show that ESMACS, using the 3-trajectory approach, is able to correctly determine binding affinities of systems that undergo large structural changes.

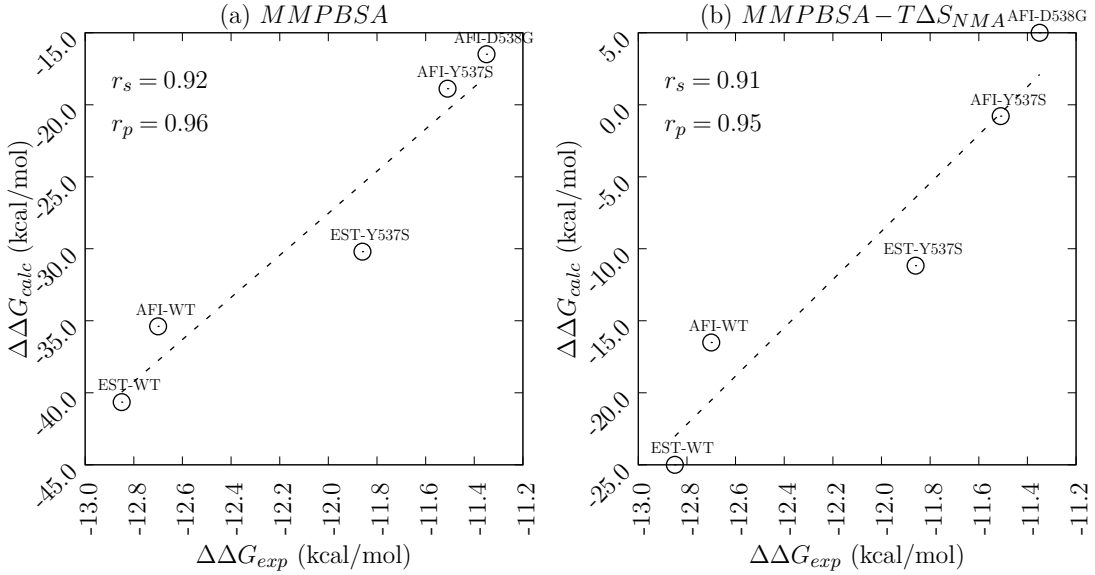


Figure 6.4: *A correlation plot between 3-trajectory ESMACS absolute binding affinities and experimental binding affinities using: (a) the MMPBSA free energy method, and (b) the MMPBSA method with the inclusion of configurational entropy obtained using normal mode analysis ($T\Delta S_{NMA}$). Error bars are removed for clarity but are no greater than ± 2 kcal/mol.*

Since the 3-trajectory approach has been implemented, this raises the question of which ER model should initially be used to generate simulation trajectories – the open or closed conformation. As we know the receptor simulation is independent of ligand binding, and sampling is solely of the receptor in solvent. We assess the 3-trajectory ESMACS binding affinities varying in the starting ER conformation. Binding affinity values are shown in Table 6.3.

When selecting either the open or closed conformation as the starting point for MD simulations, we see that the GB method, with the inclusion of configurational

Table 6.3: *Computational and experimental (ΔG_{exp}) binding affinities of estradiol, EST, and afimoxifene, AFI, to three ER receptors: WT, Y537S and D538G. The computational binding affinities are achieved using the 3-trajectory ESMACS approaches, differentiated by the starting conformation of the ER receptor prior to simulation (open and closed). The Generalised Born (GB), and Poisson-Boltzmann (PB) free energy methods were used, in addition to configuration entropy obtained from normal mode analysis (NMA). The Spearman (r_s) and Pearson (p) correlations are presented for all methods. All values are in kcal/mol.*

3-trajectory closed conformation

Ligand	Receptor	GB	PB	GB_{NM}	PB_{NM}	ΔG_{exp}
EST	WT	-29.93	-40.65	-8.87	-19.60	-12.85
EST	Y537S	-29.17	-30.20	-3.24	-9.29	-11.86
AFI	WT	-8.92	-14.12	-7.66	-12.86	-12.70
AFI	Y537S	-3.19	1.38	0.88	0.44	-11.51
AFI	D538G	4.05	-0.85	3.73	-1.18	-11.35
Spearman, r_s		0.35	0.51	0.96	0.89	
Pearson, r_p		0.59	0.71	0.98	0.94	

3-trajectory open conformation

Ligand	Receptor	GB	PB	GB_{NM}	PB_{NM}	ΔG_{exp}
EST	WT	-64.96	-61.92	-23.28	-20.25	-12.85
EST	Y537S	-54.84	-50.46	-15.18	-10.81	-11.86
AFI	WT	-43.95	-35.39	-22.07	-13.51	-12.70
AFI	Y537S	-28.86	-18.88	-11.06	-1.08	-11.51
AFI	D538G	-27.84	-16.49	-7.84	3.50	-11.35
Spearman, r_s		0.59	0.56	0.98	0.88	
Pearson, r_p		0.77	0.75	0.99	0.94	

entropy (GB_{NM}), generates better correlation and ranking coefficients. When the open conformation is selected, the GB and GB_{NM} approaches show superior performance compare to the PB variant. With regard to the closed conformation, we initially see that the GB method is worse than PB, but including the configurational entropy, GB_{NM} improves significantly and outperforms PB_{NM}. These results show us that either the open or closed conformation can be chosen as the initial starting point for MD simulations, from which binding affinities are computed. Normal mode analysis, in this case, accurately calculates the entropic changes due to conformational change in the receptor, yielding strong correlations for the full binding affinity with experimental values. Fig. 6.5 shows correlation plots for the GB variants.

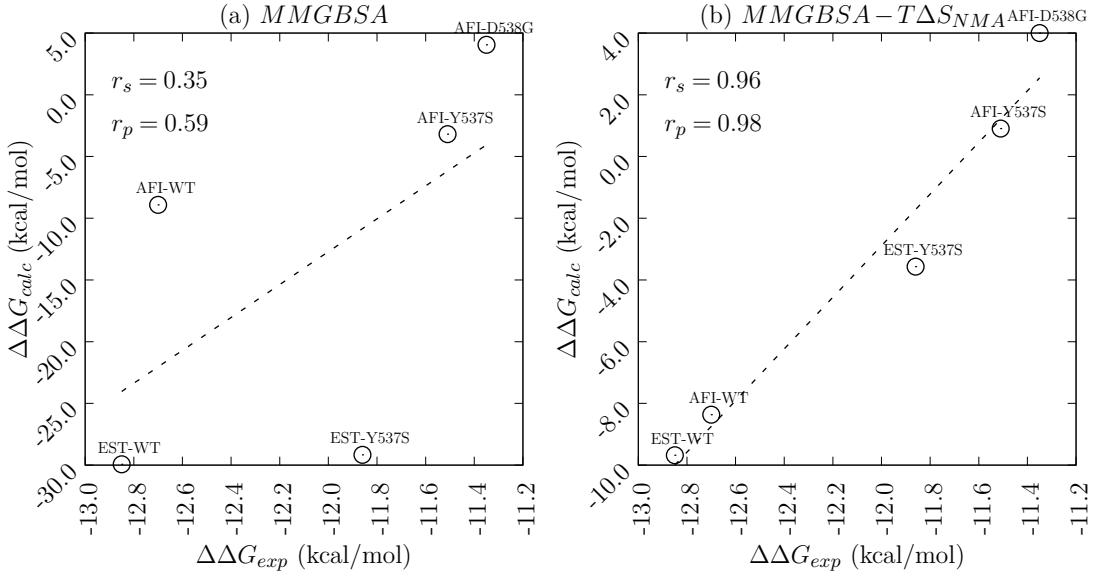


Figure 6.5: *A correlation plot between 3-trajectory ESMACS absolute binding affinities (initial receptor conformation is closed) and experimental binding affinities using: (a) the MMGBSA free energy method, and (b) the MMGBSA method with the inclusion of configurational entropy obtained using normal mode analysis ($T\Delta S_{NMA}$). Error bars are removed for clarity but are no greater than ± 2 kcal/mol.*

We have shown here that 3-trajectory ESMACS approach yields superior results compared to the single trajectory and averaged receptor approaches. This is due to the conformational changes observed upon ligand binding, and thus individual

trajectories are required to gain a more representative sampling of phase space. Normal mode analysis correctly calculates the configurational entropy as a result of the conformational change upon binding. Including the configurational entropy significantly improves the correlation and rankings. Either the open or closed receptor conformation can be selected subject to the inclusion of configurational entropy values.

6.4.1 ESMACS versus TIES relative binding affinities report a strong correlation

It is not realistic to have experimental values available each time ESMACS or TIES calculations are being performed. Clearly, if such protocols are ever to become predictive tools which clinicians can turn to with confidence, then accurate and reliable theoretical binding affinities are essential. TIES relative binding affinities were generated for the transformations described in section 6.3. By comparing TIES binding affinities with relative ESMACS binding free energy values, we are able to assess the two protocols which, as a result, gives us further confidence that ESMACS and TIES generate accurate and reliable binding affinities. Table 6.4 shows the values, correlation and ranking coefficient values.

TIES relative binding affinities show a very good agreement with the 1-trajectory ESMACS approach, exploiting the MMPBSA free energy method. This is contrary to what is seen when absolute ESMACS binding affinities are correlated against experimental values, where the 3-trajectory method shows superior correlation and ranking metrics. A similar agreement is witnessed when configurational entropy is included, albeit with a slight degradation. There is a lack of correlation between TIES and any 3-trajectory ESMACS approach. This includes when the open or closed conformation trajectories are selected.

Although the 3-trajectory ESMACS approach correctly predicts the induced-fit model that is represented by ER ligand binding, the 1-trajectory approach gener-

Table 6.4: *Relative ESMACS and TIES binding affinities for three TIES transformations, bound to three ER receptors: WT, Y537S and D538G. The relative ESMACS binding affinities are determined by finding the difference between the difference between the absolute ESMACS binding affinities, appropriate for each transform. With regard to ESMACS, the Generalised Born (GB), and Poisson-Boltzmann (PB) free energy methods were used, in addition to configuration entropy obtained from normal mode analysis (NM). The Spearman (r_s) and Pearson (r_p) correlations are presented for all end-point free energy methods, with respect to TIES binding affinities. All values are in kcal/mol.*

		1-trajectory				3-trajectory				
Receptor	Transform	GB	PB	GB_{NM}	PB_{NM}	GB	PB	GB_{NM}	PB_{NM}	$TIES$
WT	T01	2.88	1.57	1.22	-0.09	7.56	6.56	3.16	2.17	2.32
	T02	-2.77	-2.22	-2.72	-2.16	-2.71	-4.58	-3.06	-4.93	-0.61
	T03	1.24	0.04	0.28	-0.91	4.69	1.14	4.33	0.19	0.52
Y537S	T01	2.37	1.06	0.67	-0.64	0.90	2.24	-2.04	-0.70	2.25
	T02	-2.77	-2.22	-2.72	-2.16	-2.71	-4.58	-3.06	-4.93	-0.82
	T03	7.50	0.12	0.32	-1.06	-0.04	-0.79	-1.28	-2.62	0.71
D538G	T01	2.48	0.81	1.04	-0.63	3.76	4.93	0.26	1.44	2.42
	T02	-2.23	-1.88	-2.30	-1.95	7.42	8.41	7.71	8.70	-0.43
	T03	-1.63	0.46	0.89	-0.27	3.34	3.6	3.71	3.47	0.65
Spearman, r_s		0.41	0.87	0.77	0.73	0.17	0.26	0.00	0.04	
Pearson, r_p		0.64	0.93	0.88	0.85	0.42	0.51	0.02	0.19	

ated better correlations with TIES because this approach takes into account only the ligand interactions that play a direct role in binding (receptor energies are cancelled out). As a result, the 1-trajectory yields a more 'accurate' result that corresponds well with TIES. With regard to the reproducibility of the binding affinities, both approaches yield highly reproducible values. Error bars for both approaches have been removed in Fig. 6.6, for clarity. The error is no greater than ± 0.04 kcal/mol for TIES, ± 0.23 kcal/mol for ESMACS with MMPBSA, and ± 0.28 kcal/mol for ESMACS with MMPBSA−TΔS_{NMA}.

The outlier in Fig. 6.6b (T03-D538G) is as a result of an incorrect estimation of the relative binding affinity obtained from the 1-trajectory ESMACS method with the inclusion of $T\Delta S_{NMA}$. The likely reason is that the change in $T\Delta S_{NMA}$ between

TAM and TOR is considerably larger than the change in ΔG between the same ligand pair, resulting in a significant shift in binding affinity when $T\Delta S_{NMA}$ is included.

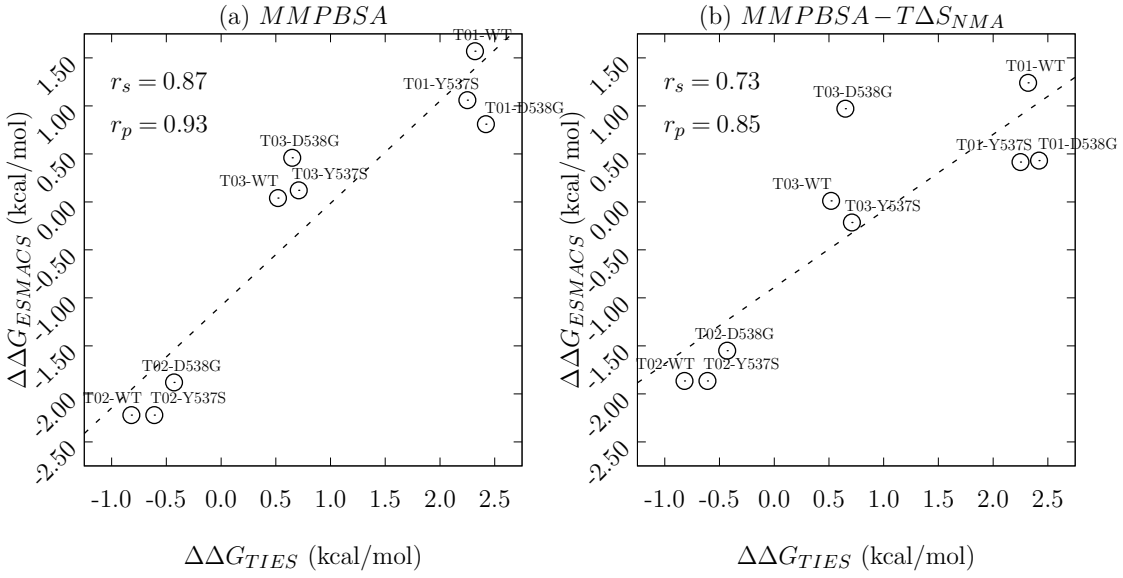


Figure 6.6: *A correlation plot between TIES relative binding affinities and 1-trajectory ESMACS relative binding affinities using: (a) the MMPBSA free energy method, and (b) the MMPBSA method with the inclusion of configurational entropy obtained using normal mode analysis ($T\Delta S_{NMA}$). Error bars for both approaches have been removed for clarity.*

6.5 Discussion

There is a need for an automated, predictive tool that can generate accurate and reliable binding affinities which clinicians can use to make rapid treatment decisions. A decision-support tool that can estimate binding affinities accurately, rapidly and reliably would help clinicians make the correct therapy choices in an efficient manner. Further, such a method would complement the rapidly improving capabilities of genome sequencing methods. Here we show that ESMACS and TIES are able to correctly estimate binding affinities for ER ligands bound to WT ER and two of the most common mutations reported in ER-positive breast cancer, Y537S and D538G.

ESMACS is a versatile protocol which has been demonstrated to work well in the ER system, and can be exploited in different ways. The 3-trajectory ESMACS approach, correctly estimates binding affinities for 5 ligands compared with experimental values. The relative ESMACS binding affinities obtained from the 1-trajectory approach show strong agreement with TIES binding affinities. This strongly validates ESMACS and TIES approaches as accurate. In scenarios where experimental values are not available, such methods would be suitable to provide reliable binding affinities. Although strong correlations and rankings have been reported in ESMACS and TIES, to further strengthen the results, additional binding affinities should be obtained.

The 3-trajectory ESMACS results correlate well with experimental binding affinities because it is able to predict binding affinities based on the induced-fit model. This is because free energy differences are obtained from individual, but ensemble-based, trajectories, where more conformations are sampled. The 1-trajectory approach on the other hand generates free energy values from only a ligand-receptor trajectory (completed as an ensemble also). As a result, most of the receptor energies are cancelled out leaving only the energy change that is directly implicated in ligand binding. This represents the lock-and-key method of binding which is not attributed to ER ligand binding. Therefore we obtain worse correlations and rankings with respect to experimental binding affinities when the 1-trajectory ESMACS approach is used. However, this is also the reason why the 1-trajectory ESMACS approach correlates well with TIES relative binding affinities. We gain a more 'accurate' binding affinity because of the fact that only key interactions in connection to ligand binding are captured.

The 3-trajectory method is limited by larger error bars. Since we obtain free energies from the complex, receptor and ligand independently, this increases the variation in free energy values. For example, the receptor free energies may vary by several kcal/mol which, when included in the bootstrapping process from which

standard errors are calculated, can contribute considerably. The 1-trajectory ESMACS approach does not have this issue for the reason that we obtain standard errors from just the ligand-receptor complex.

The open and closed receptor conformations were used to calculate the 3-trajectory binding affinities. Interestingly, we see that when selecting either conformation across all ligands, the MMGBSA method with the addition of configurational entropy report very good correlations. This example provides a further example of the fact that normal mode analysis can accurately estimate the entropic cost, in systems that have large conformational changes upon ligand binding. In previous chapters, normal mode analysis has generally shown mixed results in correlation and ranking coefficient, however this is an example where this estimation works well.

In conclusion, we have shown that the 3-trajectory ESMACS approach, with the inclusion of configurational entropy obtained by NMA, successfully predicts binding affinities for ligands bound to ER. The large conformational changes that take place upon ligand binding to ER are captured using the 3-trajectory approach, which the 1-trajectory method fails to do. This is also the first instance that we see configurational entropy significantly improving the correlation compared to MMGBSA alone, suggesting that NMA best predicts configurational entropy when systems with large conformational changes are studied. A strong correlation was reported when the relative binding affinities using 1-trajectory ESMACS was compared with TIES binding affinities. This further supports the correct calculation of absolute and relative binding affinities of ligands bound to ER.

Chapter 7

Conclusion

The question that is often raised about predictions based on computer simulation, irrespective of the scientific discipline, is if the modelling is correct. On one hand, computational models are sometimes met with skepticism by experimental scientists, and on the other, molecular modelling results have been accepted without sufficient validation. Both opinions are attributed to a lack of confidence in computational techniques that do not take into account the requirement of replica simulations to generate reproducible results. With the continuous progression of high performance computing hardware and computational techniques alike, it is now time to drive forward the development of computational tools for predictive purposes. The drug discovery domain, within the pharmaceutical industry, is an area where predictive tools are an obvious, and possibly near term, solution to stagnating productivity and rising costs. In this industry, quantifying the binding affinity of a small molecule drug to its target protein is of high interest. Two ensemble based methods, "Enhanced Sampling of Molecular dynamics with Approximation of Continuum Solvent" (ESMACS), and "Thermodynamic Integration with Enhanced Sampling" (TIES) have been described in this thesis, which have the capacity to compute accurate and reproducible binding affinities, rapidly, on an industrially relevant scale.

7.1 Summary of findings and limitations

We began with a critical analysis of ESMACS and TIES, and also compared these ensemble approaches to FEP+ (or FEP/REST), developed by Schrödinger [46, 45].

With regard to ESMACS, we generate good binding affinities for 100 ligands across all 5 systems, which vary in size and flexibility. The systems are cyclin-dependent kinase 2 (CDK2), tyrosine kinase 2 (TYK2), induced myeloid leukemia cell differentiation protein (MCL1), tyrosine-protein phosphatase non-receptor type 1 (PTP1B) and thrombin. The 1-trajectory Molecular Mechanics and the Poisson-Boltzmann Surface Area approximation (MMPBSA) yields the best correlation and ranking coefficients and also demonstrates highly reproducible results that are, in most cases, characterised by small error bars.

There are several limitations to using ESMACS. The approach is not able to sample ligand conformations that are separated by large energy barriers, which was witnessed in the bromodomain receptor [201]. This has also been identified in CDK2, where some ligands have rotamers where completely independent ESMACS calculations are required to generate binding affinities. As a result, it is not known if the starting structure, prior to molecular dynamics (MD) simulation, is correct, resulting in binding affinities that are based possibly on an incorrect ligand-receptor conformation. Enhanced sampling techniques, like metadynamics, are required to obtain a clearer understanding of ligand-receptor dynamics prior to ESMACS calculations. There were challenges in generating acceptable parameters for ligands containing a sulphonamide group. The initial parameters, give rise to incorrect partial charges on the sulphur and nitrogen atoms. These parameters were manually improved by 'dampening' the charges on these atoms. This resulted in improved binding affinity rankings and correlations for the CDK2 system. ESMACS generated binding affinities with larger error bars for charged ligands.

TIES binding affinities for the same systems that have been used in the ESMACS

study, show very good reproducibility (very small error bars), and a very high root mean square error (RMSE) and mean absolute error (MAE) coefficient, which indicates accurate and reliable relative binding affinity predictions.

The improved ligand parameters that were generated for the CDK2 system, were also employed for TIES, but resulted in very poor binding affinity values. There are several reasons why the ligand parameters seemed to work reasonably well for ESMACS, but failed to produce respectable binding affinities when TIES was employed. Firstly, and the more likely possibility, is it could be due to an anomalously high charge variations accruing in the alchemical transformation region in TIES. A less likely possibility, is that it could be a case of simply increasing the λ windows, because not enough intermediate states are sampled.

Comparing to FEP+, ESMACS correlations and rankings are significantly better than what has been published [45]. For instance, it was reported that when the Molecular Mechanics and the Generalised Born Surface Area approximation (MMGBSA) method is used to obtain binding affinities, it yields a negative correlation with experimental values. We have shown that this is not the case, and in fact we produce a reasonably good correlation. TIES results are an improvement on the FEP+ protocol, with respect to MAE and RMSE metrics. FEP+ does not take into account reproducibility; TIES demonstrates reproducibilty within carefully controlled stochastic uncertainties [42]. The FEP+ is a part of the Schrödinger suite, which is a proprietary application, includes the Optimised Potentials for Liquid Simulations (OPLS 2.1 [45]) force field, where TIES is open access. FEP+ utilises enhanced sampling in the form of the replica exchange solute tempering (REST) technique, to sample phase space that is not reachable in conventional unbiased MD. Researchers have recently implemented enhanced sampling features in TIES [73].

The objective of the study related to the PAK4 system was to improve upon binding affinities that did not produce any correlation with experimental values. As a

result we considered the effect of solvent models within our protocol, investigating the inclusion of explicit water molecules and modifying the internal dielectric parameter (ϵ_{int}) within the MMPBSA free energy method. We also assessed a method that estimates configurational entropy based on a weighted average of the solvent accessible surface area of each atom (WSASA). WSASA is an alternative to normal mode analysis (NMA) which consumes a large amounts of core hours when performing entropy calculations, within the ESMACS workflow.

Modifying the ϵ_{int} for specific ligands, in the PAK4 system shows a vast improvement in ranking and correlation. In particular, when increasing the ϵ_{int} to 4, for ligands that are comparatively less polar, we see a good correlation and ranking metric. To test this, we extended this idea to two other systems, β-site amyloid precursor protein cleaving enzyme 1 (BACE1) and reactive oxygen species 1 (ROS1), and we see a similar trend. We witness the importance of maintaining key crystal waters that are captured within the PDB crystal structure. There is a very large change in binding affinity in independent ESMACS binding affinities when crystal waters were included and excluded, respectively. The WSASA method for estimating the configurational entropy due to the binding processes has shown promising results when used for the PAK4 system. It is noteworthy that WSASA yields good correlation and rankings when ligands with aromatic variable chemical groups, in the PAK4 system, were assessed. WSASA has the benefit that it is computational undemanding (it can be completed on a regular commercial desktop) and assesses actual trajectory snapshots without the removal of explicit water molecules, as opposed to energetically minimised trajectory snapshots without solvent contributions, which required by NMA.

The inclusion of some explicit water molecules that are directly involved in ligand binding did not improve the rankings significantly. In a bid to develop a highly automated workflow to generate binding affinities, it raises the question whether manual input to include explicit water molecules is counter productive. Although

improvement in rankings have been seen when modifying ϵ_{int}, there still lacks clear and systematic criteria in the selection of ϵ_{int}. There are still issues remaining about the possible ligand parameters for the saturated and/or unsaturated rings in PAK4, and indeed, ring structures in BACE1 and ROS1. Repeating ESMACS calculations with different parameters would be beneficial in assessing this, and such a study is currently in progress with AM1-BCC [202] parameterisation method.

Binding affinity predictions are a useful asset in the clinical setting, where clinicians can be supported by accurate and reliable decision-support tools. Binding affinities using the TIES and ESMACS approach were applied to the estrogen receptor (ER) system to understand how these protocols perform for receptors that exhibit mutations – these mutations are the most common in ER-positive breast cancer. Thus, we studied the wild type (WT) ER and two mutant ERs, Y537S and D538G. Six ligands were selected of which one was the endogenous hormone estradiol, and 5 current therapies for ER-positive breast cancer.

We obtained very good results using 3-trajectory MMPBSA and MMGBSA, with the inclusion of configurational entropy obtained from NMA. This is the only instance in this dissertation that the 3-trajectory approach yields better results than 1-trajectory, but two studies have been published that report superior 3-trajectory results [100, 203]. The 3-trajectory ESMACS method is preferred in ligand-receptor systems that represent the induced-fit model, where significant conformational changes occur when the ligand is bound. This is witnessed in the ER system and is the reason for the superior binding affinity rankings when the 3-trajectory method is applied. The 1-trajectory method is better suited for the lock-and-key model that describes ligand binding. In this model, there are no large conformation changes upon ligand binding and so the majority of the free energy differences can be attributed to the interactions of the ligand with residues that comprise the binding pocket. This is also the reason why the 1-trajectory ESMACS approach yields better correlations with TIES relative binding affinities.

The inclusion of configurational entropy estimates significantly improve the correlation and ranking in the ER system, a result that has been previously reported [43]. The likely reason for this is that the conformational changes that are experienced due to ligand binding are quite large, and so NMA can more accurately estimate the entropic contribution, compared to smaller configurational changes. We also witness that selecting the open conformation of the ER gives superior results for the MMGBSA free energy method. This also requires the inclusion of an configurational entropy estimate. Relative ESMACS binding affinities (1-trajectory, MMPBSA) correlate very well with TIES binding affinities. This is a good indication that the binding affinities are correctly calculated. These methods, at least for the ER system, can be used to corroborate one another. Thus we gain even more confidence in the results which is a key characteristic for a clinical decision support tool.

7.2 Implications for future research

Ligand parameterisation is a key issue that needs to be addressed. A number of systems in this thesis have been linked with inaccurate ligand parameters, leading to poor binding affinity values. The inability to successfully generate these on a regular basis, even though the problem is an isolated issue in force fields and not directly an error in the ESMACS/TIES workflow, renders all free energy approaches unreliable. Currently, significant effort is spent on assigning correct atomic charges. With this said, considerable improvement has been made on CDK2 ligands that contain sulphonamide groups. The partial charges that were used were adopted from the CHARMM force field, so a possibility is to assess the performance of ESMACS/TIES binding affinities across widely used open-source force fields like CHARMM, GROMOS and AM1-BCC, to name three. Although this would give an indication of the performance of different force fields for particular biological system, an individual assessment of the performance of a force field needs to be made on a case-by-case basis.

FEP+ employs the REST enhanced sampling method which allows the simulation to access conformations that conventional molecular dynamics would not otherwise. This has the benefit of sampling ligand-receptor conformations that occur in reality and thus give us more accurate ligand binding affinities. TIES now has this capability by including replica exchange, and λ exchange enhanced sampling approaches [73].

ESMACS has shown good results in some cases, and modest in others. ESMACS calculations need to move away from the 'default' settings that are currently employed. This has been proved by investigating the effects of ϵ_{int} on binding affinities in PAK4, and manual manipulation of ligand partial charges in CDK2. There needs to be systematic and clear criteria when selecting important parameters when generating binding affinities. This may be specific to the polarity of the ligand, or a receptor target. However, currently there is no defined 'rule' for the selection of ϵ_{int}. Further work is required to understand when changes in ϵ_{int} should be made. These efforts should be focused on understanding the binding interfaces of the ligand and the receptor. We see that a universal ϵ_{int} value is not applicable, and so a more granular approach needs to be taken with respect to each ligand. A possible approach is to develop a model that connects the ion-ion interactions between ligands and neighbouring protein residues, and the ϵ_{int} parameter. This model would be used to select the most appropriate ϵ_{int} for the ligand-receptor simulation.

Configurational entropy is currently a limitation in the ESMACS approach. NMA requires the same amount of CPU usage, and 3 to 4 times the wall clock time to complete calculations. Compounding this, the entropy estimates are based on an energetically minimised set of trajectory snapshots that excludes solvent effects (calculations are in vacuum). The WSASA shows signs of positive results in the PAK4 system but did not improve upon the NMA results in other systems. Faster, more realistic and efficient codes such as WSASA are being implemented to speed

up this part.

It has, however, been shown that chaotic dynamics of ligand-receptor systems result in Gaussian random processes from which probabilistic binding affinities can be achieved. As such, we have high confidence in the values obtained using TIES/ESMACS supported but a tight control over errors. Both TIES and ESMACS show a high level of reproducibility that is currently not available in others methods. Coveney et al. have also developed an ensemble based method that generates absolute binding affinities from alchemical free energy techniques [73]. This follows a similar approach TIES, but with the introduction of other alchemical transformations to achieve absolute values. The approach also employs replica simulations.

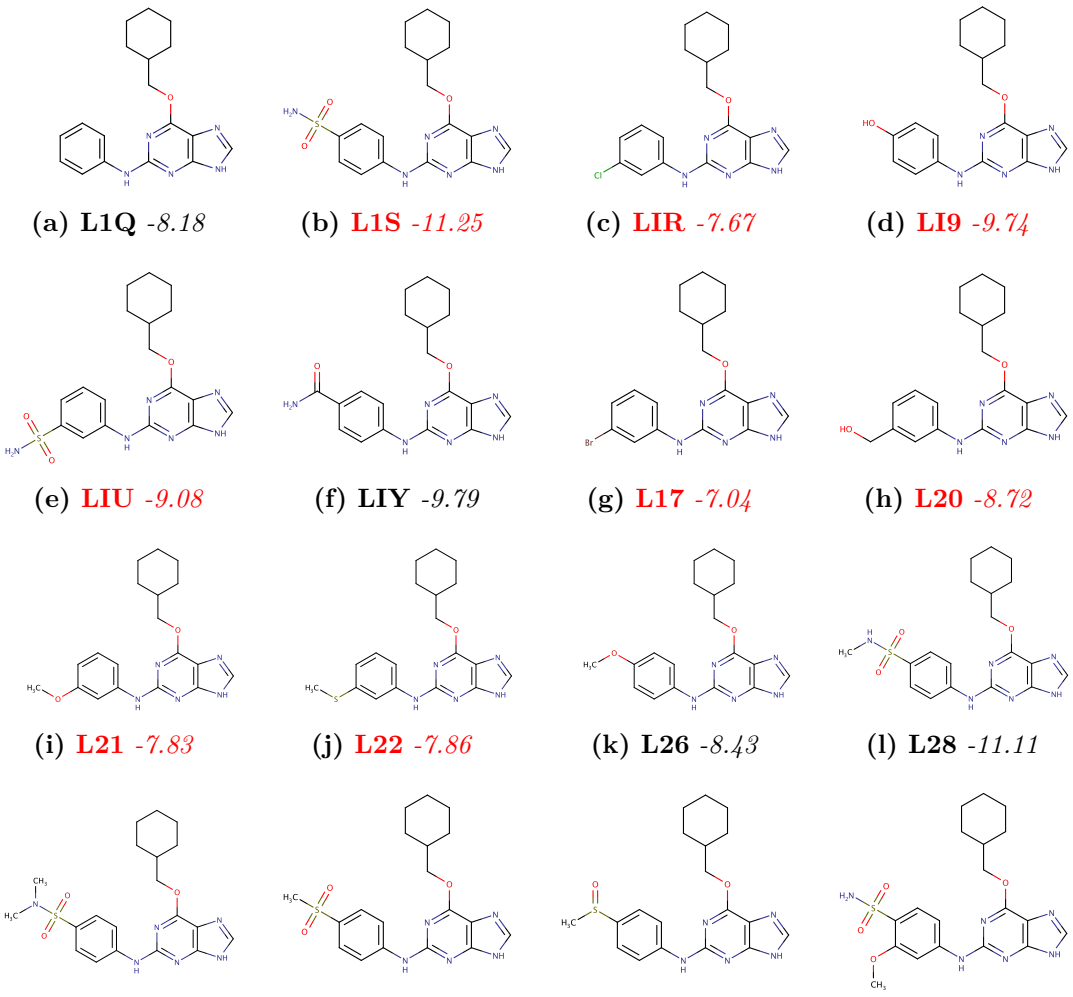
7.3 Concluding remarks

ESMACS and TIES have shown promise as tools to expedite the hit-to-lead and lead optimisation stages of drug discovery programmes and can also impact the clinical setting as decision-support tools. The protocols are distinguishable by the reproducible nature of determining binding affinities through the use of replica simulations, and a high level of automation facilitated by high-performance computers. We have seen that ESMACS and TIES are robust protocols which have been applied to a spectrum of systems containing a diverse set of ligands. The outcomes were reproducible binding affinity predictions characterised by a tight control of errors. ESMACS in particular has shown promise as a tool in the clinical setting by correctly predicting binding affinities of six ligands bound to a wild-type estrogen receptor and two mutated variants which are most commonly found in ER-positive breast cancer. With that said, both ESMACS and TIES are highly sensitive protocols which require careful manipulation of settings depending on the system of study. This was most evident when we investigated the optimal selection of ϵ_{int} in the PAK4, BACE1 and ROS1 systems. Other factors outside the

development of ESMACS and TIES include the quality of the crystal structure that is available, the reliability of the experimental binding affinity values to which a comparison is made, and the accuracy of the force field model that has been selected (particularly in the case of uncommon ligands). At this current stage, only a highly technical user is able to successfully navigate the protocol to achieve correct binding affinities. The Binding Affinity Calculator (BAC) facilitates the employment of ESMACS and TIES through an automated workflow. BAC version 2.0 now exists and a user-friendly BAC (uf-BAC) is under development, allowing non-technical users to compute binding affinities.

Appendix A

Ligand Chemical Structures



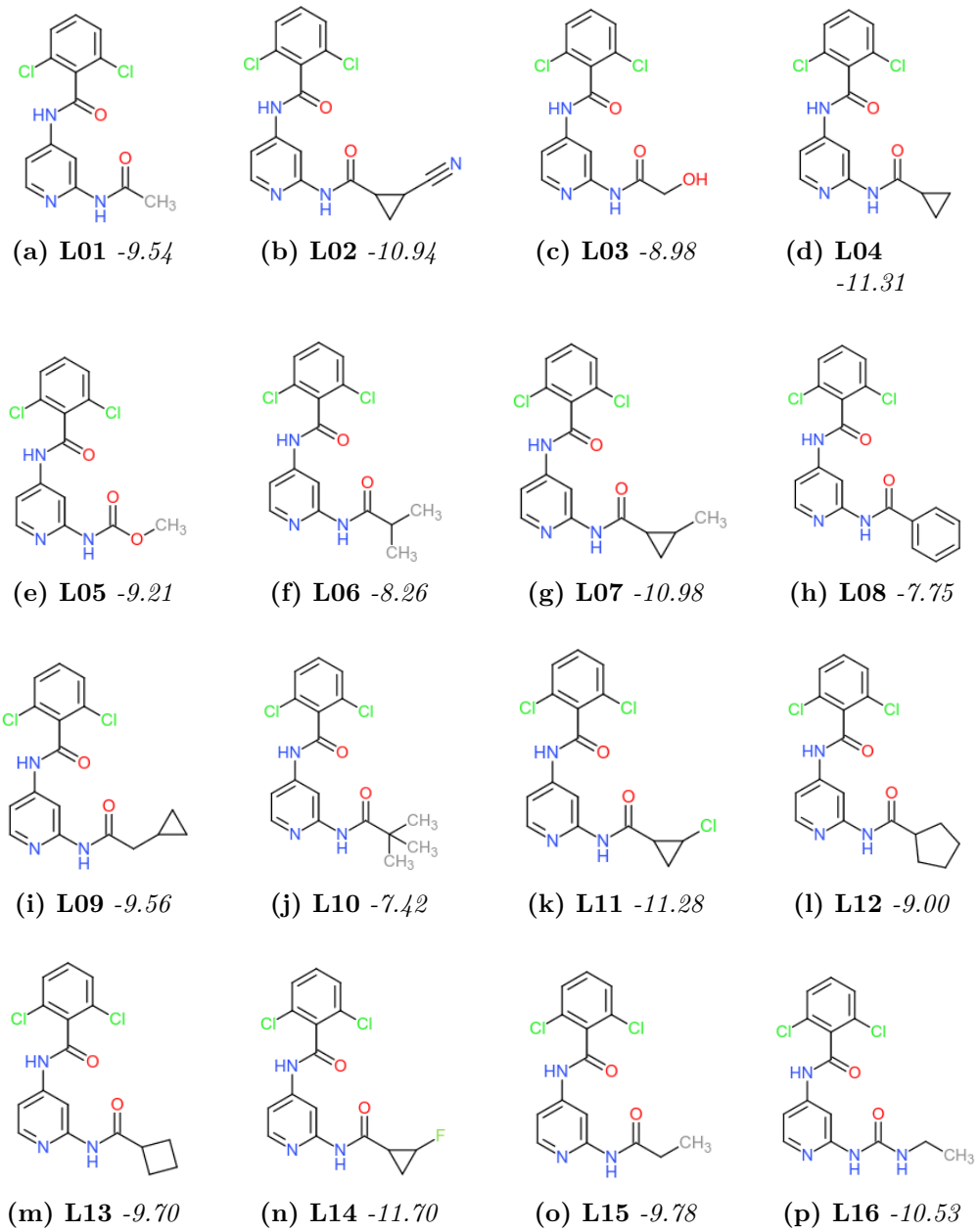
(a) **L1Q** *-8.18* (b) **L1S** *-11.25* (c) **LIR** *-7.67* (d) **LI9** *-9.74*

(e) **LIU** *-9.08* (f) **LIY** *-9.79* (g) **L17** *-7.04* (h) **L20** *-8.72*

(i) **L21** *-7.83* (j) **L22** *-7.86* (k) **L26** *-8.43* (l) **L28** *-11.11*

(m) **L29** *-9.88* (n) **L30** *-9.81* (o) **L31** *-9.54* (p) **L32** *-9.75*

Figure A.1: *Chemical structures and experimental binding affinities of 16 CDK2 inhibitors. The ligands with a meta substitution on the benzene ring exhibit rotomerism (labelled red) and thus an additional model was built. All values are reported in kcal/mol*



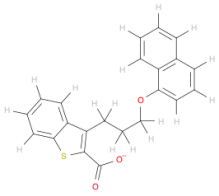
(a) **L01** *-9.54* (b) **L02** *-10.94* (c) **L03** *-8.98* (d) **L04** *-11.31*

(e) **L05** *-9.21* (f) **L06** *-8.26* (g) **L07** *-10.98* (h) **L08** *-7.75*

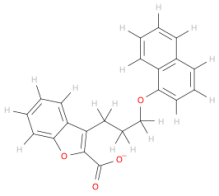
(i) **L09** *-9.56* (j) **L10** *-7.42* (k) **L11** *-11.28* (l) **L12** *-9.00*

(m) **L13** *-9.70* (n) **L14** *-11.70* (o) **L15** *-9.78* (p) **L16** *-10.53*

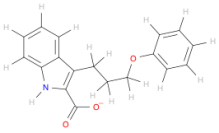
Figure A.2: *Chemical structures and associated experimental binding affinities for TY2 ligands. All values are reported in kcal/mol.*



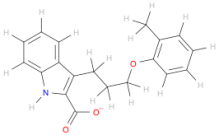
(a) **L23** *-8.83*



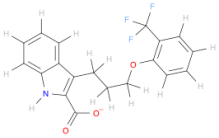
(b) **L26** *-8.24*



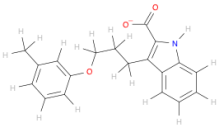
(c) **L27** *-6.12*



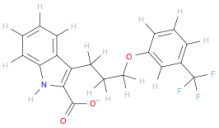
(d) **L28** *-6.62*



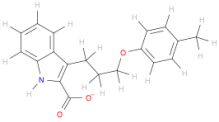
(e) **L29** *-6.94*



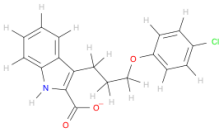
(f) **L30** *-7.85*



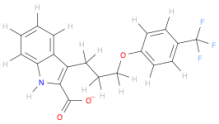
(g) **L31** *-7.92*



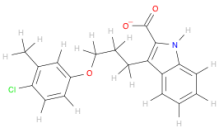
(h) **L32** *-6.58*



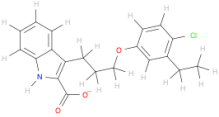
(i) **L33** *-6.88*



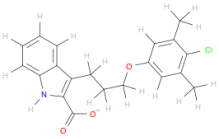
(j) **L34** *-6.87*



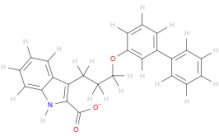
(k) **L35** *-8.81*



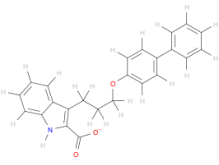
(l) **L36** *-8.18*



(m) **L37** *-8.95*



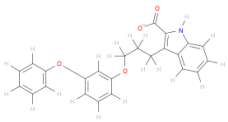
(n) **L38** *-7.02*



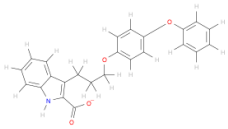
(o) **L39** *-7.03*

Figure A.3: *Chemical structures of MCL1 binding affinities and associated experimental binding affinities. All values are reported in kcal/mol.*

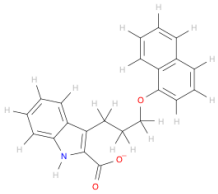
(p) **L40** *-7.25*



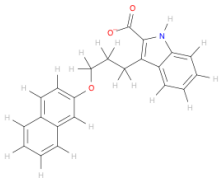
(q) **L41** *-7.13*



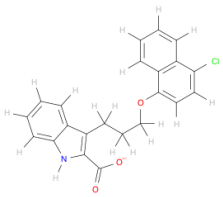
(r) **L42** *-8.90*



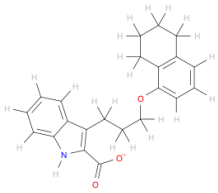
(s) **L43** *-7.03*



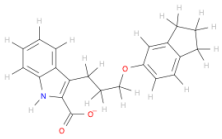
(t) **L44** *-8.67*



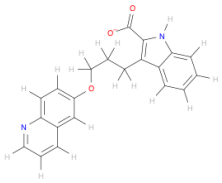
(u) **L45** *-8.95*



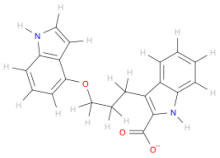
(v) **L46** *-7.60*



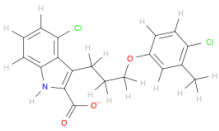
(w) **L47** *-5.78*



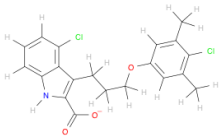
(x) **L48** *-6.66*



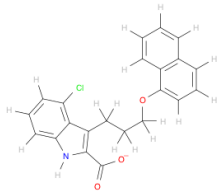
(y) **L49** *-8.36*



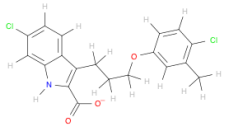
(z) **L50** *-9.33*



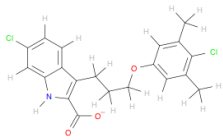
(aa) **L51** *-8.45*



(ab) **L52** *-9.23*



(ac) **L53** *-9.96*



(ad) **L54** *-9.78*

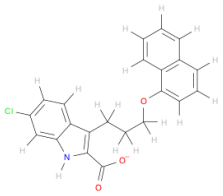
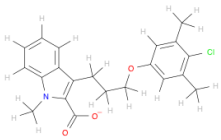
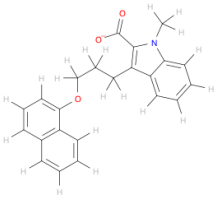


Figure A.3: *Chemical structures of MCL1 binding affinities and associated experimental binding affinities. All values are reported in kcal/mol.*

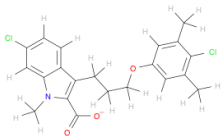
(ae) L56 *-9.26*



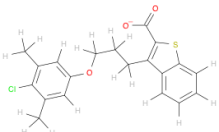
(af) L57 *-9.04*



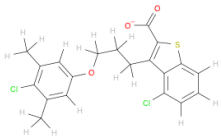
(ag) L58 *-9.41*



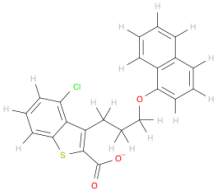
(ah) L60 *-8.92*



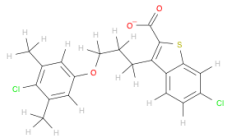
(ai) L61 *-8.08*



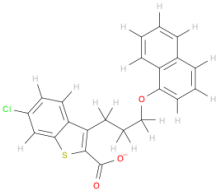
(aj) L62 *-7.96*



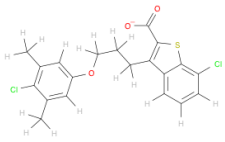
(ak) L63 *-9.06*



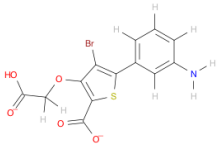
(al) L64 *-9.50*



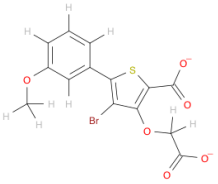
(am) L65 *-8.41*



(an) L66 *-8.43*



(ao) L67 *-7.58*



(ap) L68 *-7.69*

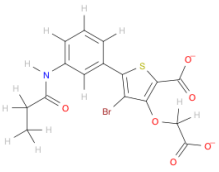
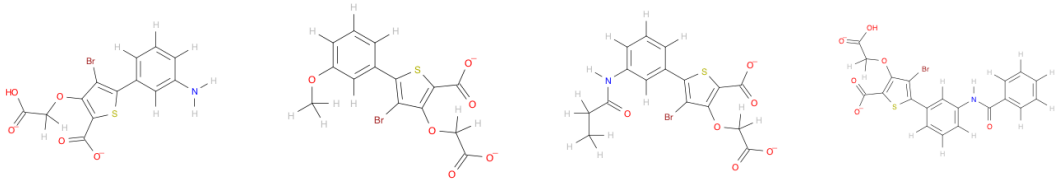


Figure A.3: *Chemical structures of MCL1 binding affinities and associated experimental binding affinities. All values are reported in kcal/mol.*

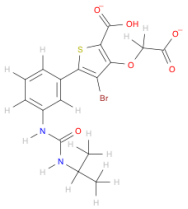


(a) L66 *-7.84*

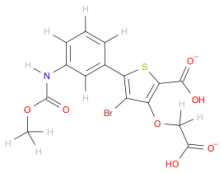
(b) L67 *-7.34*

(c) L68 *-7.74*

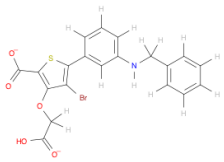
(d) L69 *-7.72*



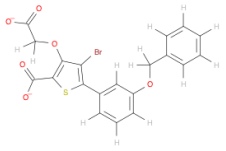
(e) L70 *-7.72*



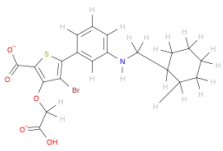
(f) L71 *-7.74*



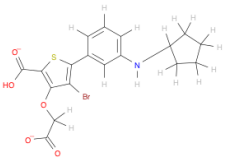
(g) L72 *-8.60*



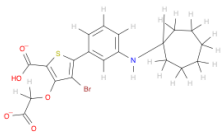
(h) L73 *-8.38*



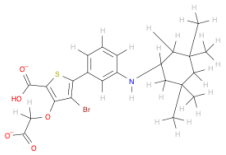
(i) L74 *-9.10*



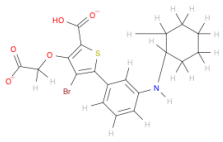
(j) L75 *-8.71*



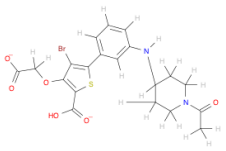
(k) L76 *-9.41*



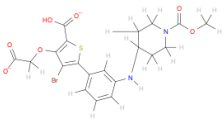
(l) L77 *-10.11*



(m) L78 *-9.08*

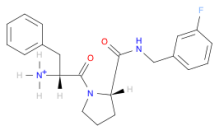


(n) L79 *-9.13*

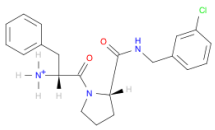


(o) L80 *-8.71*

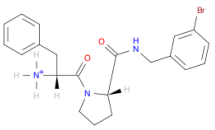
Figure A.4: *Chemical structures of PTP1B ligands and associated experimental binding affinities. All values are reported in kcal/mol.*



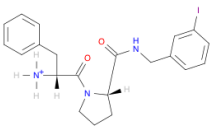
(a) L1A *-7.48*



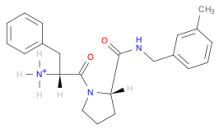
(b) L1B *-8.46*



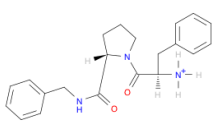
(c) L1C *-8.56*



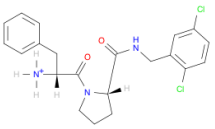
(d) L1D *-8.25*



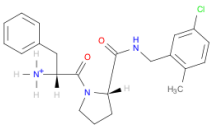
(e) L3A *-8.32*



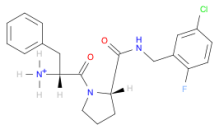
(f) L5A *-7.58*



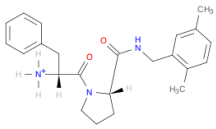
(g) L6A *-9.18*



(h) L6B *-8.89*



(i) L6E *-8.91*



(j) L7A *-8.22*

Figure A.5: *Chemical structures of thrombin ligands and associated experimental binding affinities. All values are reported in kcal/mol.*

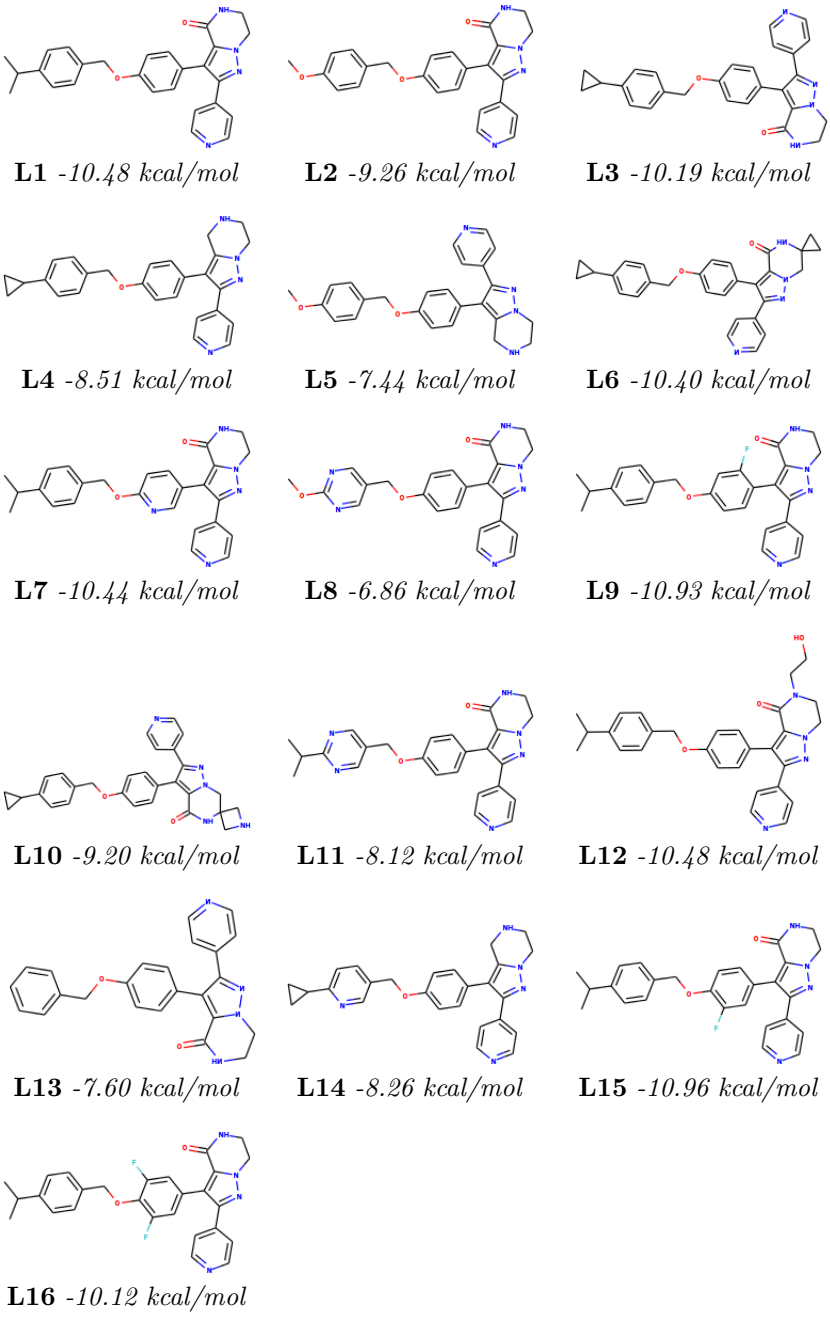
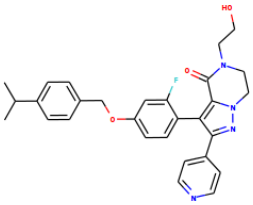
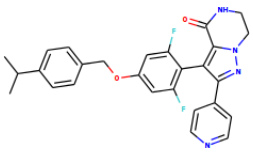


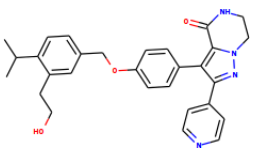
Figure A.6: *Chemical Structures and binding affinities of ROS1 Ligands*



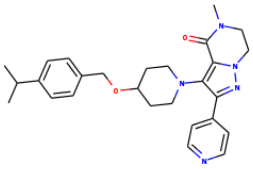
L17 *-10.97 kcal/mol*



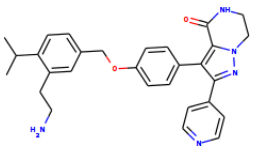
L18 *-8.57 kcal/mol*



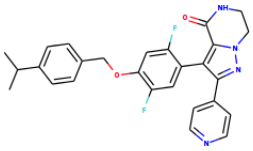
L19 *-10.93 kcal/mol*



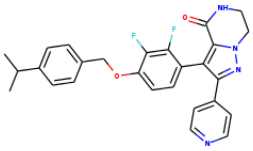
L20 *-9.49 kcal/mol*



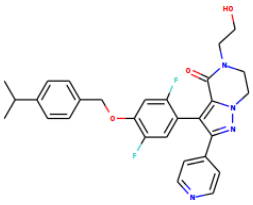
L21 *-10.80 kcal/mol*



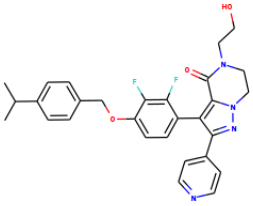
L22 *-11.21 kcal/mol*



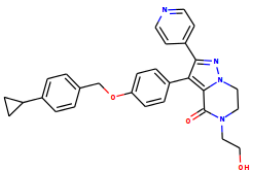
L23 *-9.60 kcal/mol*



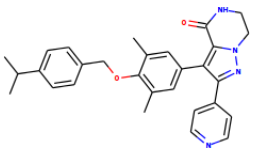
L24 *-11.48 kcal/mol*



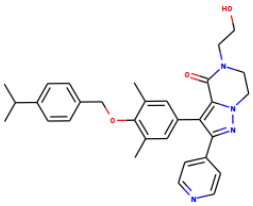
L25 *-9.58 kcal/mol*



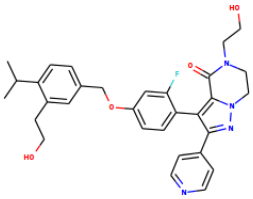
L26 *-9.95 kcal/mol*



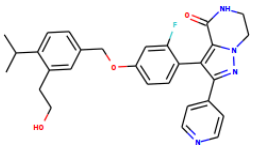
L27 *-6.86 kcal/mol*



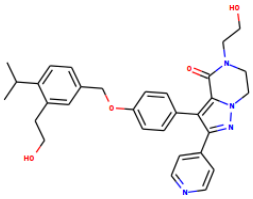
L28 *-6.86 kcal/mol*



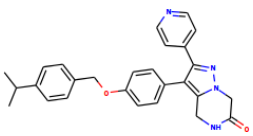
L29 *-11.13 kcal/mol*



L30 *-11.29 kcal/mol*

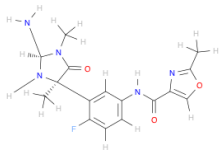


L31 *-10.62 kcal/mol*

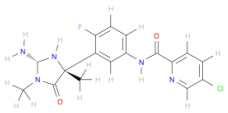


L32 *-8.40 kcal/mol*

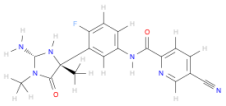
Figure A.6: *Cont... Chemical Structures and binding affinities of ROS1 Ligands*



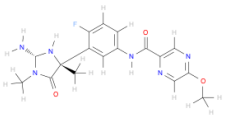
L01 *-7.86*



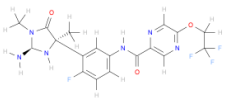
L04 *-7.60*



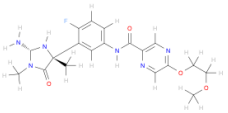
L07 *-5.08*



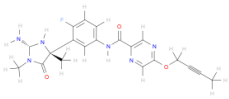
L10 *-10.01*



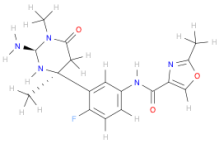
L13 *-9.50*



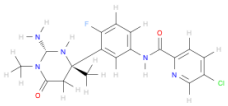
L16 *-7.14*



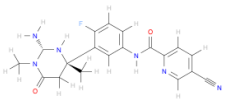
L19 *-8.92*



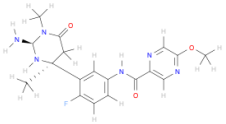
L02 *-9.19*



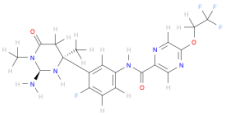
L05 *-6.89*



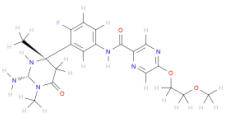
L08 *-9.57*



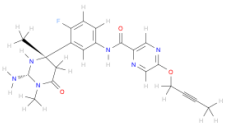
L11 *-9.16*



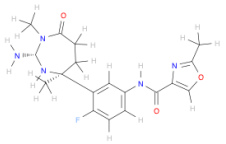
L14 *-7.05*



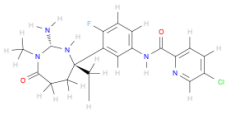
L17 *-7.09*



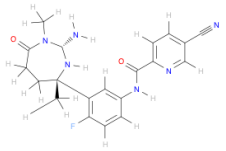
L21 *-7.20*



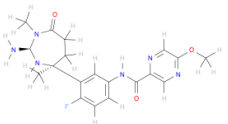
L03 *-2.93*



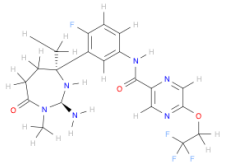
L06 *-10.47*



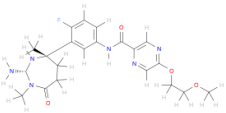
L09 *-10.41*



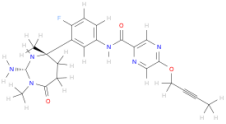
L12 *-6.84*



L15 *-7,65*



L18 *-5.49*



L20 *-9.61*

Figure A.7: *Chemical structures of BACE1 ligands and associated experimental binding affinities. All values are reported in kcal/mol.*

Appendix B

ESMACS and TIES Binding Affinity Tables

Table B.1: *Binding free energies using the 1-trajectory ESMACS approach for CDK2 ligands bound to the CDK2 receptor. GB/PB is the free energy method using the Generalised Born or Poisson Boltzmann approximation. GB/PB_{NM} is the above method with the inclusion of configurational entropy term. All values are in kcal/mol.*

	1-trajectory				
Ligand	GB	PB	GB_{NM}	PB_{NM}	ΔG_{exp}
L1Q	-35.50	-29.61	-15.82	-9.92	-8.18
L1S	-39.24	-36.50	-16.48	-13.73	-11.25
L1R	-41.97	-35.96	-22.48	-16.47	-7.67
L1R-R	-41.98	-35.30	-22.48	-15.81	-7.67
LI9	-37.27	-32.88	-17.72	-13.34	-9.74
LIU	-40.32	-36.86	-16.36	-12.91	-9.08
LIU-R	-41.50	-38.27	-17.88	-14.65	-9.08
LIY	-42.32	-39.35	-21.28	-18.31	-9.79
L17	-36.57	-31.83	-17.25	-12.53	-7.04
L17-R	-37.28	-31.96	-17.29	-11.97	-7.04
L20	-36.67	-31.41	-14.49	-9.23	-8.72
L20-R	-42.67	-36.97	-19.71	-14.01	-8.72
L21	-34.85	-30.81	-14.30	-10.25	-7.83
L21-R	-35.95	-30.30	-15.21	-9.57	-7.83
L22	-39.60	-35.36	-19.16	-14.92	-7.86
L22-R	-41.99	-36.54	-21.69	-16.23	-7.86
L26	-38.01	-32.55	-16.92	-11.46	-8.43
L28	-44.36	-38.76	-22.91	-17.32	-11.11
L29	-38.11	-33.95	-15.49	-11.39	-9.88
L30	-40.58	-37.27	-20.01	-16.69	-9.81
L31	-40.97	-36.10	-20.49	-15.62	-9.54
L32	-38.41	-37.52	-15.93	-15.04	-9.75
L32-R	-40.47	-35.83	-18.38	-13.72	-9.75

Table B.2: *Binding free energies using the 2-trajectory ESMACS approach for CDK2 ligands bound to the CDK2 receptor. GB/PB is the free energy method using the Generalised Born or Poisson Boltzmann approximation. GB/PB_{NM} is the above method with the inclusion of configurational entropy term. All values are in kcal/mol.*

	2-trajectory				
Ligand	GB	PB	GB_{NM}	PB_{NM}	ΔG_{exp}
L1Q	-335.94	-359.52	-322.49	-346.06	-8.18
L1S	1717.47	1727.46	1715.21	1725.21	-11.25
L1R	85.03	111.26	112.69	138.92	-7.67
L1R-R	81.00	105.31	108.93	133.24	-7.67
LI9	-337.18	-363.76	-325.59	-352.18	-9.74
LIU	7.17	45.54	39.86	78.22	-9.08
LIU-R	10.11	49.73	41.12	80.75	-9.08
LIY	78.43	100.30	107.33	129.20	-9.79
L17	-328.83	-354.13	-315.90	-341.21	-7.04
L17-R	-333.98	-362.02	-320.37	-348.41	-7.04
L20	-332.05	-356.67	-317.76	-342.38	-8.72
L20-R	-337.75	-362.39	-321.53	-346.17	-8.72
L21	-325.94	-354.28	-313.21	-341.54	-7.83
L21-R	-333.93	-354.34	-320.56	-340.98	-7.83
L22	85.86	111.07	113.99	139.20	-7.86
L22-R	80.29	103.75	109.63	133.09	-7.86
L26	-330.17	-354.74	-314.96	-339.53	-8.43
L28	78.66	101.47	108.08	130.89	-11.11
L29	-332.62	-360.82	-317.44	-345.70	-9.88
L30	83.32	105.11	111.11	132.90	-9.81
L31	85.41	108.40	114.02	137.01	-9.54
L32	14.81	55.07	44.47	84.72	-9.75
L32-R	14.19	56.25	43.64	85.71	-9.75

Table B.3: *Binding free energies using the 3-trajectory ESMACS approach for CDK2 ligands bound to the CDK2 receptor. GB/PB is the free energy method using the Generalised Born or Poisson Boltzmann approximation. GB/PB_{NM} is the above method with the inclusion of configurational entropy term. All values are in kcal/mol.*

	3-trajectory				
Ligand	GB	PB	GB_{NM}	PB_{NM}	ΔG_{exp}
L1Q	-335.69	-359.19	-322.33	-345.83	-8.18
L1S	1718.22	1728.13	1715.80	1725.71	-11.25
L1R	88.49	114.01	115.94	141.46	-7.67
L1R-R	85.47	109.00	113.14	136.68	-7.67
LI9	-336.63	-363.17	-325.11	-351.66	-9.74
LIU	7.64	45.94	40.19	78.49	-9.08
LIU-R	11.54	51.59	42.59	82.64	-9.08
LIY	82.22	103.27	111.06	132.11	-9.79
L17	-327.53	-352.75	-314.86	-340.09	-7.04
L17-R	-332.94	-360.94	-319.56	-347.56	-7.04
L20	-331.52	-356.06	-317.48	-342.02	-8.72
L20-R	-336.69	-361.50	-320.68	-345.49	-8.72
L21	-325.81	-354.01	-313.11	-341.30	-7.83
L21-R	-332.87	-353.17	-319.72	-340.03	-7.83
L22	59.25	83.80	87.16	111.71	-7.86
L22-R	54.52	77.32	83.52	106.31	-7.86
L26	-330.13	-354.58	-314.98	-339.42	-8.43
L28	83.27	105.21	112.57	134.51	-11.11
L29	-355.17	-379.11	-340.72	-364.72	-9.88
L30	86.80	107.74	114.50	135.44	-9.81
L31	89.74	111.86	118.29	140.41	-9.54
L32	15.54	55.85	45.29	85.59	-9.75
L32-R	16.22	58.16	45.51	87.46	-9.75

Table B.4: *TIES relative binding free energies for CDK2 ligand pairs. 'Initial' and 'Final' indicate the starting end end ligands of the respective TIES transformations. ΔG^{com}_{alch} is the free energy of the alchemical transformation bound to the receptor, and ΔG^{aq}_{alch} is the free energy of the alchemical transformation in aqueous solution. $\Delta\Delta G_{calc}$ and $\Delta\Delta G_{exp}$ are the calculated and experimental relative binding affinities, respectively. All values are in kcal/mol.*

Name	Initial	Final	ΔG^{com}_{alch}	ΔG^{aq}_{alch}	$\Delta\Delta G_{calc}$	$\Delta\Delta G_{exp}$
T00	L28	L29	76.60	76.73	0.13	-1.23
T01	L1S	L28	2.47	2.41	-0.07	-0.14
T02	L1S	L30	4.65	4.31	-0.33	-1.44
T03	LI9	L26	4.97	4.21	-0.75	-1.31
T04	L20	L21	-9.33	-11.69	-2.35	-0.89
T05	LIY	L31	-8.37	-8.45	-0.07	-0.25
T06	LIY	L30	-13.89	-13.53	0.36	0.02
T07	L1Q	L1R	1.36	0.01	-1.35	-0.51
T08	L1Q	L1S	-2.59	-1.96	0.63	3.07
T09	L1Q	LIU	-45.38	-40.51	4.87	0.90
T10	L1Q	LIY	-47.52	-45.50	2.01	1.61
T11	L1Q	L22	-4.32	-2.97	1.35	-0.32
T12	L1Q	L28	-57.11	-55.13	1.98	2.93
T13	L1Q	L30	-62.01	-59.62	2.38	1.63
T14	L1Q	L31	-52.80	-51.50	1.30	1.36
T15	L1Q	L32	-76.95	-75.97	0.98	1.57
T16	L1Q	LI9	14.08	15.86	1.78	1.56
T17	L1Q	L20	-20.33	-18.48	1.85	0.54
T18	L1Q	L21	-29.50	-30.11	-0.61	-0.35
T19	L1Q	L26	18.72	19.51	0.79	0.25
T20	L1Q	L29	-55.91	-54.06	1.85	1.70
T21	L1Q	L17	-14.85	-15.33	-0.48	-1.14

Table B.5: *Binding free energies using the 1-trajectory ESMACS approach for TYK2 ligands bound to the TYK2 receptor. GB/PB is the free energy method using the Generalised Born or Poisson Boltzmann approximation. GB/PB_{NM} is the above method with the inclusion of configurational entropy term. All values are in kcal/mol.*

1-trajectory					
Ligand	GB	PB	GB_{NM}	PB_{NM}	ΔG_{exp}
L01	-30.77	-12.80	-30.07	-12.09	-9.54
L02	-36.25	-16.93	-35.18	-15.86	-10.94
L03	-29.42	-10.46	-29.19	-10.23	-8.98
L04	-33.48	-16.95	-31.98	-15.45	-11.31
L05	-32.72	-15.34	-30.33	-12.95	-9.21
L06	-29.47	-11.23	-29.44	-11.19	-8.26
L07	-34.33	-16.13	-32.65	-14.45	-10.98
L08	-33.76	-13.10	-31.41	-10.75	-7.75
L09	-32.72	-14.04	-31.26	-12.58	-9.56
L10	-26.67	-7.01	-29.37	-9.71	-7.42
L11	-35.23	-17.96	-33.81	-16.54	-11.28
L12	-32.47	-13.19	-31.66	-12.38	-9
L13	-31.47	-12.82	-30.96	-12.30	-9.7
L14	-33.85	-15.84	-32.49	-14.48	-11.7
L15	-31.25	-13.64	-30.60	-12.99	-9.78
L16	-37.39	-19.84	-33.81	-16.26	-10.53

Table B.6: *Binding free energies using the 2-trajectory ESMACS approach for TYK2 ligands bound to the TYK2 receptor. GB/PB is the free energy method using the Generalised Born or Poisson Boltzmann approximation. GB/PB_{NM} is the above method with the inclusion of configurational entropy term. All values are in kcal/mol.*

	2-trajectory				
Ligand	GB	PB	GB_{NM}	PB_{NM}	ΔG_{exp}
L01	-26.94	-10.20	-29.08	-12.33	-9.54
L02	-35.93	-17.22	-36.30	-17.59	-10.94
L03	-34.61	-15.67	-32.09	-13.14	-8.98
L04	-35.85	-18.08	-34.26	-16.49	-11.31
L05	-27.77	-11.02	-25.38	-8.62	-9.21
L06	-31.50	-12.84	-31.48	-12.82	-8.26
L07	-33.53	-18.29	-31.41	-16.17	-10.98
L08	-27.68	-7.05	-26.22	-5.58	-7.75
L09	-36.66	-15.76	-31.95	-11.05	-9.56
L10	-28.86	-11.02	-31.52	-13.67	-7.42
L11	-34.63	-17.65	-35.33	-18.34	-11.28
L12	-32.40	-12.69	-31.04	-11.33	-9
L13	-30.17	-12.13	-29.09	-11.05	-9.7
L14	-31.77	-13.25	-31.28	-12.76	-11.7
L15	-34.87	-15.46	-33.81	-14.39	-9.78
L16	-38.07	-18.96	-33.98	-14.87	-10.53

Table B.7: *Binding free energies using the 3-trajectory ESMACS approach for TYK2 ligands bound to the TYK2 receptor. GB/PB is the free energy method using the Generalised Born or Poisson Boltzmann approximation. GB/PB_{NM} is the above method with the inclusion of configurational entropy term. All values are in kcal/mol.*

3-trajectory					
Ligand	GB	PB	GB_{NM}	PB_{NM}	ΔG_{exp}
L01	-24.44	-53.93	-545.97	-574.73	-9.54
L02	-31.96	-66.00	-422.00	-454.28	-10.94
L03	-32.20	-61.68	-430.07	-458.77	-8.98
L04	-32.40	-63.84	-407.34	-437.75	-11.31
L05	-25.32	-55.81	-518.26	-548.35	-9.21
L06	-28.87	-61.28	-406.14	-437.45	-8.26
L07	-28.56	-64.58	-475.81	-510.79	-10.98
L08	-24.96	-55.74	-418.91	-449.08	-7.75
L09	-34.64	-65.30	-492.75	-522.62	-9.56
L10	-26.03	-60.61	-392.02	-425.40	-7.42
L11	-30.81	-65.37	-440.01	-473.55	-11.28
L12	-30.19	-63.68	-384.17	-416.93	-9
L13	-27.89	-59.75	-394.81	-426.01	-9.7
L14	-27.77	-59.64	-459.19	-490.21	-11.7
L15	-32.29	-61.93	-443.81	-472.55	-9.78
L16	-36.01	-65.76	-631.39	-661.59	-10.53

Table B.8: *Binding free energies using the 1- 2- and 3-trajectory ESMACS approach for MCL1 ligands bound to the MCL1 receptor. GB/PB is the free energy method using the Generalised Born or Poisson Boltzmann approximation. The configurational entropy term was not estimated for this system. All values are in kcal/mol.*

	1-trajectory		2-trajectory		3-trajectory		
Ligand	*GB*	*PB*	*GB*	*PB*	*GB*	*PB*	ΔG_{exp}
L23	-53.29	-45.17	-52.90	-44.49	-52.05	-43.61	-8.83
L26	-50.63	-43.00	-51.87	-42.99	-50.71	-41.74	-8.24
L27	-42.74	-36.54	-46.67	-42.04	-45.99	-41.27	-6.12
L28	-43.24	-37.31	-43.62	-38.95	-41.73	-36.83	-6.62
L29	-45.36	-39.00	-44.99	-37.99	-42.56	-35.61	-6.94
L30	-44.84	-38.78	-46.90	-40.98	-46.13	-40.18	-7.85
L31	-43.71	-38.28	-39.44	-34.10	-38.58	-33.21	-7.92
L32	-43.32	-37.66	-38.93	-32.79	-37.68	-31.55	-6.58
L33	-42.81	-36.74	-47.26	-42.72	-46.25	-41.68	-6.88
L34	-42.48	-36.81	-44.50	-38.20	-43.18	-36.81	-6.87
L35	-47.36	-41.63	-45.82	-39.49	-44.83	-38.47	-8.81
L36	-49.66	-43.74	-45.08	-39.49	-43.79	-38.13	-8.18
L37	-47.68	-43.42	-46.00	-43.71	-45.15	-42.77	-8.95
L38	-51.11	-43.42	-53.40	-44.95	-50.72	-42.23	-7.02
L39	-38.17	-33.81	-38.48	-34.55	-36.42	-32.40	-7.03
L40	-46.83	-40.51	-48.32	-43.05	-45.82	-40.49	-7.25
L41	-44.53	-39.48	-46.62	-41.01	-45.04	-39.48	-7.13
L42	-51.96	-44.11	-52.87	-45.89	-51.66	-44.69	-8.9
L43	-47.32	-40.36	-47.06	-39.15	-45.92	-37.99	-7.03
L44	-51.97	-44.72	-52.14	-44.25	-50.98	-43.09	-8.67
L45	-54.06	-48.01	-55.09	-50.75	-54.23	-49.85	-8.95
L46	-50.47	-43.52	-49.30	-41.10	-48.70	-40.46	-7.6
L47	-43.04	-37.12	-44.98	-38.49	-43.67	-37.11	-5.78
L48	-43.31	-36.02	-44.61	-36.63	-42.52	-34.52	-6.66
L49	-42.32	-38.39	-44.98	-40.46	-42.50	-37.89	-8.36
L50	-46.55	-43.25	-49.45	-44.83	-46.83	-42.12	-9.33
L51	-48.77	-41.59	-50.24	-43.30	-47.66	-40.51	-8.45
L52	-51.12	-45.25	-49.48	-42.77	-49.04	-42.27	-9.23
L53	-54.03	-49.28	-51.40	-45.47	-50.74	-44.86	-9.96
L54	-55.17	-47.28	-53.87	-46.55	-53.01	-45.73	-9.78
L56	-52.67	-47.03	-52.71	-47.67	-52.11	-47.10	-9.26
L57	-54.01	-44.87	-52.42	-44.56	-51.29	-43.46	-9.04
L58	-56.22	-50.33	-58.17	-53.20	-57.47	-52.66	-9.41
L60	-51.66	-47.02	-53.98	-48.91	-53.61	-48.55	-8.92
L61	-48.03	-44.29	-46.05	-41.54	-44.38	-39.79	-8.08
L62	-49.73	-42.71	-48.29	-40.78	-46.79	-39.09	-7.96
L63	-53.78	-49.26	-51.19	-46.01	-50.96	-45.87	-9.06
L64	-54.24	-46.27	-51.62	-43.51	-50.87	-42.82	-9.5
L65	-54.10	-49.11	-53.31	-48.91	-52.81	-48.52	-8.41
L66	-55.67	-47.53	-54.92	-47.89	-54.21	-47.17	-8.43
L67	-50.84	-46.65	-50.45	-45.59	-49.02	-44.24	-7.58
L68	-51.38	-43.71	-50.82	-43.24	-48.75	-41.18	-7.69

Table B.9: *Binding free energies using the 1- 2- and 3-trajectory ESMACS approach for PTP1B ligands bound to the PTP1B receptor. GB/PB is the free energy method using the Generalised Born or Poisson Boltzmann approximation. The configurational entropy term was not estimated for this system. All values are in kcal/mol.*

	1-trajectory		2-trajectory		3-trajectory		
Ligand	GB	PB	GB	PB	GB	PB	ΔG_{exp}
L66	-55.65	-45.83	-52.30	-40.60	-51.64	-40.17	-7.84
L67	-58.52	-48.17	-60.42	-49.13	-59.79	-48.70	-7.34
L68	-58.13	-48.64	-58.75	-51.68	-57.97	-51.15	-7.74
L69	-57.64	-47.06	-59.99	-50.35	-59.27	-49.82	-7.72
L70	-59.31	-49.28	-54.96	-44.11	-53.94	-43.22	-7.72
L71	-58.16	-48.25	-57.30	-47.56	-56.50	-47.00	-7.74
L72	-57.54	-44.63	-56.05	-40.36	-55.19	-39.69	-8.60
L73	-54.55	-43.03	-54.69	-41.18	-54.11	-40.80	-8.38
L74	-59.73	-47.84	-62.09	-50.79	-61.29	-50.21	-9.10
L75	-57.39	-46.94	-61.76	-52.31	-61.10	-51.82	-8.71
L76	-58.05	-46.88	-58.91	-49.33	-57.48	-48.12	-9.41
L77	-48.53	-46.17	-50.23	-48.57	-49.01	-47.37	-10.12
L78	-57.50	-45.88	-51.48	-39.18	-49.71	-37.66	-9.08
L79	-59.78	-45.11	-59.30	-46.28	-57.82	-44.97	-9.13
L80	-59.97	-46.30	-62.22	-48.56	-60.90	-47.44	-8.71

Table B.10: *Binding free energies using the 1- 2- and 3-trajectory ESMACS approach for thrombin ligands bound to thethrombin receptor. GB/PB is the free energy method using the Generalised Born or Poisson Boltzmann approximation. The configurational entropy term was not estimated for this system. All values are in kcal/mol.*

	1-trajectory		2-trajectory		3-trajectory		
Ligand	GB	PB	GB	PB	GB	PB	ΔG_{exp}
L1A	-37.24	-32.48	-40.06	-34.41	-38.08	-32.16	-7.48
L1B	-41.58	-35.39	-41.10	-35.25	-39.20	-33.00	-8.46
L1C	-43.38	-36.88	-45.63	-38.85	-43.83	-36.63	-8.56
L1D	-40.28	-35.97	-43.98	-41.03	-41.35	-38.25	-8.25
L3A	-39.57	-35.56	-37.72	-35.04	-35.18	-31.97	-8.32
L3B	-43.52	-38.43	-41.68	-37.34	-38.92	-34.67	-7.86
L5A	-37.11	-32.50	-37.42	-33.76	-34.94	-31.35	-7.58
L6A	-42.35	-38.65	-42.77	-35.91	-39.99	-32.88	-9.18
L6B	-43.37	-37.92	-42.25	-36.13	-38.39	-32.67	-8.89
L6E	-40.45	-35.76	-36.89	-31.83	-34.38	-29.08	-8.91
L7A	-40.37	-36.16	-39.71	-36.14	-35.58	-32.55	-8.22

Table B.11: *Spearman rank (r_s) and Pearson rank (r_p) correlation for all trajectory ESMACS approaches, with the inclusion of crystal waters. Metrics are report with and without ligand L01 to highlight the reliance of this ligand to the initial correlation that is observed.*

		1-traj		2-traj		3-traj	
		r_s	r_p	r_s	r_p	r_s	r_p
All	GB	0.16	0.40	0.23	0.48	0.24	0.49
	PB	0.23	0.48	0.09	0.30	0.08	0.29
	$GB-T\Delta S$	0.25	0.50	0.37	0.61	0.41	0.64
	$PB-T\Delta S$	0.37	0.61	0.18	0.42	0.19	0.43
no L01	GB	0.02	-0.14	0.04	0.20	0.04	0.21
	PB	0.03	0.17	0.01	0.08	0.01	0.07
	$GB-T\Delta S$	0.06	0.24	0.21	0.46	0.26	0.51
	$PB-T\Delta S$	0.22	0.47	0.06	0.25	0.08	0.28

Table B.12: *Binding free energies obtained using the 1-trajectory ESMACS approach for PAK4 ligands with varying internal dielectric values ϵ_{int}. Values are shown for the Poisson-Boltzmann free energy method ($MMPBSA$) and the same method with the inclusion of configurational entropy, estimated using normal mode analysis $MMPBSA_{NM}$. All values are in kcal/mol.*

	$MMPBSA$			$MMPBSA_{NM}$			
Ligand	$\epsilon_{int} = 1$	$\epsilon_{int} = 4$	$\epsilon_{int} = 10$	$\epsilon_{int} = 1$	$\epsilon_{int} = 4$	$\epsilon_{int} = 10$	ΔG_{exp}
L01	-39.53	-57.96	-61.64	-16.53	-34.96	-38.64	-11.70
L03	-31.95	-41.42	-43.32	-9.99	-19.46	-21.36	-11.66
L04	-32.26	-42.11	-44.08	-9.63	-19.48	-21.45	-10.62
L18	-30.43	-39.78	-41.64	-8.53	-17.88	-19.74	-11.26
L19	-29.99	-39.39	-41.27	-7.30	-16.70	-18.58	-8.38
L20	-31.58	-40.01	-41.70	-7.37	-15.80	-17.49	-8.33
L22	-32.21	-39.88	-41.42	-9.83	-17.50	-19.04	-9.17
L23	-33.31	-41.13	-42.70	-12.36	-20.18	-21.75	-9.89
L24	-31.39	-39.68	-41.34	-9.45	-17.74	-19.40	-8.07
L25	-32.35	-41.37	-43.17	-11.45	-20.47	-22.27	-7.75
L27	-33.56	-40.18	-41.51	-14.27	-20.89	-22.22	-9.35
L29	-32.81	-40.68	-42.25	-12.53	-20.40	-21.97	-8.50
L30	-32.28	-40.78	-42.49	-11.48	-19.98	-21.69	-8.80
L31	-33.71	-41.86	-43.49	-13.28	-21.43	-23.06	-8.39

Table B.13: *Binding free energies obtained using the 1-trajectory ESMACS approach for PAK4 ligands with a varying number of explicit water molecules included in the free energy calculation. Values are shown for the Poisson-Boltzmann ($MMPBSA$) and Generalised Born free energy method ($MMGBSA$) without configurational entropy. All values are in kcal/mol.*

1-trajectory $MMPBSA$

Ligand	10	20	30	40	50	60	70	0
L01	-40.13	-40.08	-39.36	-39.53	-39.77	-39.88	-39.94	-39.17
L03	-29.33	-29.36	-28.77	-28.92	-29.14	-29.25	-29.30	-31.89
L04	-31.70	-31.97	-31.37	-31.58	-31.80	-31.95	-31.98	-32.18
L18	-30.48	-29.94	-30.02	-30.26	-30.38	-30.45	-30.47	-30.37
L19	-27.71	-26.15	-25.41	-25.31	-25.46	-25.58	-25.65	-29.91
L20	-28.16	-25.68	-24.97	-24.99	-25.16	-25.26	-25.28	-31.53
L22	-27.82	-25.21	-24.61	-24.68	-24.86	-24.95	-24.98	-32.15
L23	-29.00	-29.62	-29.48	-29.78	-30.03	-30.14	-30.22	-33.22

1-trajectory $MMGBSA$

Ligand	10	20	30	40	50	60	70	0
L01	-47.93	-45.49	-43.83	-42.23	-40.87	-39.87	-39.17	-52.59
L03	-33.90	-29.56	-26.17	-23.21	-20.86	-19.22	-17.94	-39.97
L04	-35.74	-33.59	-31.56	-29.67	-28.16	-27.16	-26.36	-40.19
L19	-35.88	-33.14	-31.12	-29.41	-28.24	-27.39	-26.71	-39.43
L20	-32.53	-28.46	-25.72	-23.46	-21.69	-20.39	-19.40	-40.00
L22	-32.14	-26.81	-24.11	-21.86	-20.09	-18.85	-17.90	-40.82
L23	-32.60	-27.57	-25.25	-23.31	-21.78	-20.76	-19.96	-36.84
L18	-34.43	-34.02	-32.90	-31.71	-30.75	-30.14	-29.64	-41.18

Table B.14: *Binding free energies using the 1-trajectory ESMACS approach for ROS1 ligands bound to the ROS1 receptor. $MMPBSA$ is the free energy method using the Poisson Boltzmann approximation. $MMPBSA_{NM}$ is the above method with the inclusion of configurational entropy term. All values are in kcal/mol.*

	1-trajectory		
Ligand	$MMPBSA$	$MMPBSA_{NM}$	ΔG_{exp}
L01	-41.04	-41.04	-10.48
L02	-38.13	-38.13	-9.26
L03	-39.75	-39.75	-10.19
L04	-41.17	-41.17	-8.51
L05	-38.43	-38.43	-7.44
L06	-41.60	-41.60	-10.40
L07	-43.05	-45.65	-10.44
L08	-33.15	-33.15	-6.86
L09	-43.09	-43.09	-10.93
L10	-39.03	-39.03	-9.20
L11	-37.97	-37.97	-8.12
L12	-44.43	-47.09	-10.48
L13	-34.50	-34.50	-7.60
L14	-39.51	-39.51	-8.26
L15	-42.22	-42.22	-10.96
L16	-44.30	-44.30	-10.12
L17	-45.57	-48.46	-10.97
L18	-41.20	-41.20	-8.57
L19	-42.03	-45.00	-10.93
L20	-48.97	-48.97	-9.49
L21	-41.99	-41.99	-10.80
L22	-44.13	-44.13	-11.21
L23	-44.02	-44.02	-9.56
L24	-47.76	-50.77	-11.48
L25	-46.25	-49.28	-9.58
L26	-42.16	-42.16	-9.95
L27	-43.06	-29.53	-6.86
L28	-45.74	-29.49	-6.86
L29	-45.29	-48.96	-11.13
L30	-43.12	-46.35	-11.29
L31	-43.60	-46.95	-10.62
L32	-42.79	-42.79	-8.40

Table B.15: *Binding free energies using the 1-trajectory ESMACS approach for BACE1 ligands bound to the BACE1 receptor. $MMPBSA$ is the free energy method using the Poisson Boltzmann approximation. $MMPBSA_{NM}$ is the above method with the inclusion of configurational entropy term. All values are in kcal/mol.*

	1-trajectory		
Ligand	$MMPBSA$	$MMPBSA_{NM}$	ΔG_{exp}
L01	-39.81	-32.82	-3.38
L02	-40.00	-40.00	-4.75
L03	-37.55	-37.55	-3.97
L04	-43.45	-39.46	-3.69
L05	-43.57	-43.57	-5.02
L06	-41.34	-41.34	-4.33
L07	-42.01	-30.82	-3.60
L08	-42.49	-42.49	-5.10
L09	-40.87	-40.87	-4.34
L10	-43.10	-30.30	-3.62
L11	-42.79	-42.79	-4.94
L12	-41.40	-41.40	-4.02
L13	-44.17	-37.13	-3.65
L14	-45.07	-45.07	-5.18
L15	-40.83	-40.83	-4.06
L16	-47.06	-39.79	-3.24
L17	-47.52	-47.52	-5.05
L18	-44.45	-44.45	-3.95
L19	-44.29	-33.20	-4.49
L20	-42.87	-42.87	-4.82
L21	-46.55	-46.55	-5.31

Bibliography

[1] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw. Biomolecular simulation: A computational microscope for molecular biology. *Ann. Rev. Biophys.*, 41(1):429–452, 2012.

[2] R .P. Feynman, R. B. Leighton, and M. L. Sands. *The Feynman Lectures on Physics*. Addison-Wesley, 1963.

[3] A. Todd, R. J. Anderson, and P. W. Groundwater. Rational drug design - identifying and characterising a target. *Pharm. J.*, 283:19–20, 2009.

[4] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Rev. Drug Discov.*, 9(3):203–214, 2010.

[5] F. Pammolli, L. Magazzini, and M. Riccaboni. The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discov.*, 10(6):428–438, 2011.

[6] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen. Innovation in the pharmaceutical industry: New estimates of r & d costs. *J. Health Econ.*, 47:20–33, 2016.

[7] R. W. Hansen. The pharmaceutical development process: estimates of current development costs and times and the effects of regulatory changes. *Iss.*

Pharm. Econ., pages 151–187, 1979.

[8] J. A. DiMasi, R. W. Hansen, H. G. Grabowski, and L. Lasagna. Cost of innovated in the pharmaceutical industry. *J. Health Econ.*, 10:107–142, 1991.

[9] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski. The price of innovation: New estimates of drug development costs. *J. Health Econ.*, 22(2):151–185, 2003.

[10] J. Drews and others. Drug discovery: a historical perspective. *Science*, 287(5460):1960–19644, 2000.

[11] P. G. Wyatt, I. H. Gilbert, K. D. Read, and A. H. Fairlamb. Target validation: linking target and chemical properties to desired product profile. *Curr. Top. Med. Chem.*, 11(10):1275–83, 2011.

[12] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

[13] R. Nuzzo. How scientists fool themselves - and how they can stop. *Nature News*, 526(7572):182–185, 2015.

[14] F. Prinz, T. Schlange, and K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Rev. Drug Disc.*, 10(9):712–713, 2011.

[15] M. Furber, F. Narjes, and J. Steel, editors. *The Handbook of Medicinal Chemistry*. The Royal Society of Chemistry, 2013.

[16] S. Wan, B. Knapp, D. Wright, C. Deane, and P. V. Coveney. Rapid, precise, and reproducible prediction of Peptide-MHC binding affinities from molecular dynamics that correlate well with experiment. *J. Chem. Theory. Comput.*, 11:3346–3356, 2015.

[17] S.K. Sadiq, D.W. Wright, O.A. Kenway, and P. V. Coveney. Accurate ensemble molecular dynamics binding free energy of multidrug-resistance hiv-1 protease. *J. Chem. Inf. Model.*, 50:890–905, 2010.

[18] T.D. Bunney, W. Shunzhou, N. Thiyagarajan, L. Sutto, S. V. Williams, P. Ashford, H. Koss, M. A. Knowles, F. L. Gervasio, P. V. Coveney, and M. Katan. The effect of mutations on drug sensitivity and kinase activity of fibroblast growth factor receptors: a combined experimental and theoretical study. *EBioMedicine*, 2:194–204, 2015.

[19] S. Wan and P. V. Coveney. Rapid and accurate ranking of binding affinities of epidermal growth factor receptor sequences with selected lung cancer drugs. *J. R. Soc. Interface*, 8:1114–1127, 2011.

[20] Computation Biomedicine – A Centre of Excellence in Computational Biomedicine. http://www.compbiomed.eu/. Accessed: 19 January 2018.

[21] H. P. Rang, M. M. Dale, and J. M. Ritter. *Pharmacology.* Churchill Livingstone, 5th edition, 2003.

[22] D. T. Haynie. *Biological Thermodynamics.* Cambridge University Press, 2 edition, 2008.

[23] L. Styer, J. M. Berg, and J. L. Tymoczko. *Biochemistry.* W. H .Freeman & Co. Ltd, 5th edition, 2002.

[24] L. Michaelis and M . L. Menten. Die kinetik der invertinwirkung. *Biochemische Zeitschrift*, 45:333–369, 1913.

[25] H. Lineweaver and D. Burk. The determination of enzyme dissociation constants. *J. Am. Chem. Soc.*, 56(3):658–666, 1934.

[26] T. D. Pollard. A guide to simple and informative binding assays. *Mol. Biol. Cell*, 21(23):4061–4067, 2010.

[27] M. M. Pierce, C. S. Raman, and B. T. Nall. Isothermal titration calorimetry of protein-protein interactions. *Methods*, 19(2):213–221, 1999.

[28] A. M. Rossi and C. W. Taylor. Analysis of protein-ligand interactions by fluorescence polarization. *Nat. Protocols*, 6:365–387, 2011.

[29] J. C. Owicki. Fluorescence polarization and anisotropy in high throughput screening: Perspectives and primer. *J. Biomol. Screen*, 5:297–306, 2000.

[30] K. E. Sapsford, L. Berti, and I.L. Medintz. Materials for fluorescence resonance energy transfer analysis: Beyond traditional donor-acceptor combinations. *Angewandte Chemie*, 45:4562–4589, 2006.

[31] A. R. Clapp, Medintz I. L., et al. Fluorescence resonance energy transfer between quantum dot donors and dye-labeled protein acceptors. *J. Am. Chem. Soc.*, 126:301–310, 2004.

[32] C. Yung-Chi and W. H. Prusoff. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition of an enzymatic reaction. *Biochem. Pharmacology*, 22(23):3099 – 3108, 1973.

[33] R. L. Rich and D. G. Myszka. A new platform for routine biomolecular interaction analysis. *J. Mol. Recognit.*, 14:223–228, 2001.

[34] R. Ghai, R. J. Falconer, and B. M. Collins. Applications of isothermal titration calorimetry in pure and applied research-survey of the literature from 2010. *J. Mol. Recognit.*, 25:32–52, 2012.

[35] J. B. Chaires. Calorimetry and thermodynamics in drug design. *Annu. Rev. Biophys.*, 37:135–151, 2008.

[36] Harding S. E. and B. Chowdhry. *Protein-Ligand Interactions: Hydrodynamics and Calorimetry*. Oxford University Press, 2001.

[37] P. L. Kastritis, I. H. Moal, and others. A structure-based benchmark for protein-protein binding affinity. *Protein Sci.*, 20:482–491, 2011.

[38] T. Arai, M. Yatabe, and others. A fluorescence polarization-based assay for the identification and evaluation of calmodulin antagonists. *Anal. Biochem*, 405:147–152, 2010.

[39] F. Jensen. *Introduction to Computational Chemistry.* Wiley, 2nd edition, 2011.

[40] A. R. Leach. *Molecular Modelling: Principles and Applications.* Prentice Hall, 2nd edition, 2001.

[41] P. V. Coveney and S. Wan. On the calculation of equilibrium thermodynamic properties from molecular dynamics. *Phys. Chem. Chem. Phys.*, 2016.

[42] A. Bhati, S. Wan, D. W. Wright, and P. V. Coveney. Rapid, accurate, precise and reliable relative free energy prediction using ensemble based thermodynamic integration. *J. Chem. Theor. Comp*, 13:210–222, 2016.

[43] D. Wright, B. Hall, O. Kenway, S. Jha, and P. V. Coveney. Computing Clinically Relevant Binding Free Energies of HIV-1 Protease Inhibitors. *J. Chem. Theory Comput.*, 10:1228–1241, 2014.

[44] M. Aldeghi, A. Heifetz, et al. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.*, 7:207–218, 2015.

[45] L. Wang, Y. Wu, et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.*, 137:2695–703, 2015.

[46] L. Wang, Y. Deng, et al. Modeling local structural rearrangements using FEP/REST: Application to relative binding affinity predictions of CDK2 inhibitors. *J. Chem. Theor. Comp.*, 9:1282–1293, 2013.

[47] N. A. Metropolis, A. W. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[48] Clark, S. Monte Carlo or Molecular Dynamics. http://cmt.dur.ac.uk/sjc/. Accessed: 31 January 2018.

[49] L. Verlet. Computer "experiments" on classical fluids. I. thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, 159:98, 1967.

[50] W. F. Van Gunsteren and H. J. C. Berendsen. A leap-frog algorithm for stochastic dynamics. *Molec. Sim.*, 1(3):173–185, 1988.

[51] S. Toxvaerd. Molecular dynamics at constant temperature and pressure. *Phys. Rev. E.*, 47(1):343–350, 1993.

[52] H. C. Andersen. Molecular dynamics simulations at constant temperature and/or pressure. *J. Chem. Phys.*, 72(4):2384–2393, 1980.

[53] H. J. C. Berendsen, J. P. M. Postma, W. F. Gunsteren, A. Di Nola, and J. R. Haak. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[54] S. A. Adelman and J. D. Doll. Generalized Langevin equation approach for atom/solid-surface scattering: General formulation for classical scattering of harmonic solids. *J. Chem. Phys.*, 64:2375–2388, 1976.

[55] S. A. Adelman. Generalized langevin theory for many-body problems in chemical dynamics: General formulation and the equivalent harmonic chain representation. *J. Chem. Phys.*, 71:4471–4486, 1979.

[56] C. L. Brooks III, B. M. Pettitt, and M. Karplus. Structural and energetic effects of truncating long ranged interactions in ionic and polar fluids. *J. Chem. Phys.*, 85:5897–5908, 1985.

[57] P. P. Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annal. Phys.*, 369:1521–3889, 1921.

[58] T. Darden, D. York, and L. Pedersen. Particle mesh ewald: An $N \cdot log(N)$ method for ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.

[59] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comp. Phys.*, 23(3):327–341, 1977.

[60] V. Krautler, W. F. van Gunsteren, and P. H. Hunenberger. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J.Comp. Chem.*, 22:501–508, 2001.

[61] S. Miyamoto and P. A. Kollman. An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.*, 13:952–962, 1992.

[62] O. Guvench and A. D. MacKerell. Comparison of protein force fields for molecular dynamics simulations. *Methods Molec. Biol.*, 443:63–88, 2008.

[63] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *JACS*, 117(19):5179–5197, 1995.

[64] A. MacKerell, D. Bashford, M. Bellot, R. Dunbrack, J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. Lau, C. Mattos, S. Michnick, T. Ngo, D. Nguyen, B. Prodhom, W. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-Atom Empirical Po-

tential for Molecular Modeling and Dynamics Studies of Proteins. *J. of Phys. Chem.*, 102:3586–3616, 1998.

[65] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comp. Chem.*, 25(13):1656–1676, 2004.

[66] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *JACS*, 118(45):11225–11236, 1996.

[67] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general AMBER force field. *J. Comput. Chem.*, 25:1157–74, 2004.

[68] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, 31:671–90, 2010.

[69] T. Straatsma, H. Berendsen, and J. Postma. Free energy of hydrophobic hydration: A molecular dynamics study of nobel gases in water. *J. Chem. Phys.*, 85:6720–6727, 1986.

[70] T. Straatsma and H. Berendsen. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulation. *J. Chem. Phys.*, 89:5876–5886, 1988.

[71] R. Zwanzig. High-temperature equation of state by a perturbation method in non polar gases. *J. Chem. Phys.*, 22:1420–1426, 1954.

[72] D. A. Pearlman. Evaluating the molecular mechanics Poisson-Boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase. *J. Med. Chem*, pages 7796–7807, 2005.

[73] A. P. Bhati, S. Wan, Y. Hu, B. Sherborne, and P. V. Coveney. Uncertainty Quantification in Alchemical Free Energy Methods. *J. Chem. Theory Comput.*, 14(6):2867–2880, 2018.

[74] X. Du, Y. Li, Y. Xia, et al. Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. *Int. J. Molec. Sci.*, 17(2):144–168, 2016.

[75] D. Green and R. Leach. Comuter-aided molecular design under the SWOT-light. *J. Comput. Aided Mol. Des.*, 26:51–56, 2012.

[76] G. M. Morris, R. Huey, W. Lindstrom, and others. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 30:2785–2791, 2009.

[77] G. Jones, P. Willet, R. C. Glen, and others. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267:727–748, 1997.

[78] S. Mukherjee, T. E. Balius, and R. C. Rizzo. Docking validation resources: Protein family and ligand flexibility experiments. *J. Chem. Inf. Model.*, 50:1986–2000, 2010.

[79] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.*, 15:411–428, 2001.

[80] M. Rarey, B. Kramer, T. Lengauer, and G. Kleve. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261:470–489, 1996.

[81] R. A. Friesner, R. B. Murphy, M. P. Repasky, et al. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.*, 49:470–489, 2006.

[82] J. Aqvist, C. Medina, and J-E. Samuelsson. A new method for predicting binding affinity in computer-aided drug design. *Prot. Engin.*, 7:385–391, 1994.

[83] A. Onufriev, D. Bashford, and A. Case. Modification of the generalized Born model suitable for macromolecules. *J. Phys. CHem.. B*, 104:3712–3720, 2000.

[84] B. R. Brooks, D. Janezic, and M. Karplus. Harmonic analysis of large systems. I. methodology. *J. Comput. Chem.*, 16:1522–1542, 1995.

[85] C. Chang, W. Wei, and M. K. Gilson. Evaluating the accuracy of the quasi-harmonic approximation. *J. Chem. Theor. Comput. Phys.*, 1:1017–1028, 2005.

[86] J. Wang and T. Hou. Develop and test a solvent accessible surface area-based model in conformational entropy calculations. *J. Chem. Inf. Model.*, 52(5):1199–1212, 2012.

[87] S. Sadiq, D. Wright, S. Watson, S. Zasada, I. Stoica, and P. V. Coveney. Automated molecular simulation based binding affinity calculator for ligand-bound HIV-1 protease. *J. Chem. Inf. Model.*, 48:1909–1919, 2008.

[88] M. J. Frisch, G. W. Trucks, J. A. Pople, et al. Gaussian 03, revision c.02.

[89] Case D. A., Kollman P. A., et al. Amber 14.

[90] J. A. Maier, C. Martinez, K. Kasavajhala, et al. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–3713, 2015.

[91] N. Homeyer, A. Horn, H. Lanig, and H. Stricht. AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J. Mol. Mod.*, 12:281–289, 2006.

[92] P. Mark and L. Nilsson. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *J. Phys. Chem.*, 105:9954–9960, 2001.

[93] J. C. Phillips et al. Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 26:1781–1802, 2005.

[94] D. A. Case, T. E. Cheatham, et al. The AMBER biomolecular simulation programs. *J. of Comp. Chem.*, 26(16):1668–1688, 2005.

[95] D. A. Pearlman. A comparison of alternative approaches to free energy calculations. *J. Phys. Chem*, pages 1487–1493, 1994.

[96] M. Zacharias, T. P. Straatsma, and J. A. McCammon. Separation shifted scaling, a new scaling method for lennard jones interactions in thermodynamic integration. *J. Chem. Phys.*, 100:9025–9031, 1994.

[97] T. C. Beutler, A. E. Mark, P. R. van Schaik, R. C.and Gerber, and W. F. van Gunsteren. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem. Phys. Lett.*, 222:529–539, 1994.

[98] T. S. Lee, Y. Hu, B. Sherborne, Z. Guo, and D. M.. York. Toward Fast and Accurate Binding Affinity Prediction with pmemdGTI: An Efficient Implementation of GPU-Accelerated Thermodynamic Integration. *J. Chem. Theor. Comput.*, 13(7):3077–3084, 2017.

[99] D. Groen, A. P. Bhati, J. Suter, J. Hetherington, S. J. Zasada, and P. V. Coveney. Fabsim: Facilitating computational research through automation

on large-scale and distributed e-infrastructures. *Comp. Phys. Comm.*, 207:375 – 385, 2016.

[100] S. Wan, A. P. Bhati, S. J. Zasada, I. Wall, D. Green, and Coveney P. V. Rapid and reliable binding affinity prediction for analysis of bromodomain inhibitors : a computational study. *J. Chem. Theor. Comp.*, pages 1–30, 2016.

[101] R. Highfield. Supercomputer bid to create the first truly personalised medicine. https://blog.sciencemuseum.org.uk/supercomputer-bid-to-create-the-first-truly-personalised-medicine/, 2016. Accessed: 29 November 2016.

[102] Gauss Supercomputing Centre. SuperMUC Enables Major Finding in Personalized Medicine. https://www.hpcwire.com/off-the-wire/supermuc-enables-major-finding-personalized-medicine/, 2016. Accessed: 20 August 2017.

[103] STFC Hartree Centre. STFC Hartree Centre HPC Resources. http://yukon.dl.ac.uk:8080/wiki/site/admin/resources.html. Accessed: 18 January 2018.

[104] ARHER Hardware. http://www.archer.ac.uk/about-archer/hardware/. Accessed: 18 January 2018.

[105] SuperMUC Petascale System Specification. https://www.lrz.de/services/compute/supermuc/systemdescription/. Accessed: 18 January 2018.

[106] C. J. Sherr. G1 phase progression: Cycling on cue. *Cell*, 79:551 – 555, 1994.

[107] A. A. Russo, P. D. Jeffrey, and N. P. Pavletich. Structural basis of cyclin-dependent kinase activation by phosphorylation. *Nat. Struct. Biol.*, 3:696–700, 1996.

[108] I. Hardcastle, C. Arris, et al. N2-substituted O6-cyclohexylmethylguanine derivatives: potent inhibitors of cyclin-dependent kinases 1 and 2. *J. Med. Chem.*, 47:3710–22, 2004.

[109] K. Ghoreschi, A. Laurence, and J. J. O'Shea. Janus kinases in immune cell signaling. *Immunol. Rev.*, 228(1):273–287, 2009.

[110] A.O. Chua, R. Chizzonite, B.B. Desai, T.P. Truitt, P. Nunes, L.J. Minetti, R.R. Warrier, D.H. Presky, J.F. Levine, M.K. Gately, and U. Gubler. Expression cloning of a human IL-12 receptor component: A new member of the cytokine receptor superfamily with strong homology to gp130. *J. Immunol.*, 153(1):128–136, 1994.

[111] C. Parham et al. A receptor for the heterodimeric cytokine IL-23 is composed of IL-12Rβ1 and a novel cytokine receptor subunit, IL-23R. *J. Immunol.*, 168(11):5699–5708, 2002.

[112] K. Boniface, B. Blom, Y. Liu, and R. De Waal Malefyt. From interleukin-23 to T-helper 17 cells: human T-helper cell differentiation revisited. *Immunol. Rev.*, 226(1):132–146, 2008.

[113] C. Langrish, B. McKenzie, N. Wilson, R. De Waal Malefyt, R. Kastelein, and D. Cua. IL-12 and IL-23: master regulators of innate and adaptive immunity. *Immunol. Rev.*, 202(1):96–105, 2004.

[114] R. Nair et al. Genomewide scan reveals association of psoriasis with IL-23 and NF-κB pathways. *Nat. Genet.*, 41(2):199–204, 2009.

[115] P. Ahern, A. Izcue, K. Maloy, and F. Powrie. The interleukin-23 axis in intestinal inflammation. *Immunol. Rev.*, 226(1):147–159, 2008.

[116] J. M. Adams and S. Cory. The Bcl-2 apoptotic switch in cancer development and therapy. *Oncogene*, 26(9):1324–1337, 2007.

[117] S. Willis et al. Apoptosis initiated when BH3 ligands engage multiple Bcl-2 homologs, not bax or bak. *Science*, 315(5813):856–859, 2007.

[118] N. Danial and S. Korsmeyer. Cell death: Critical control points. *Cell*, 116(2):205–219, 2004.

[119] R. Beroukhim et al. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463:899, 2010.

[120] G. Wei et al. Chemical genomics identifies small-molecule MCL1 repressors and BCL-xL as a predictor of MCL1 dependency. *Cancer Cell*, 21(4):547–562, 2012.

[121] H. Zhang, S. Guttikonda, L. Roberts, T. Uziel, D. Semizarov, S. W. Elmore, J. D. Leverson, and L. T. Lam. Mcl-1 is critical for survival in a subgroup of non-small-cell lung cancer cell lines. *Oncogene*, 30:1963–1968, 2010.

[122] J. Danilewicz et al. Design of selective thrombin inhibitors based on the (R)-Phe-Pro-Arg sequence. *J. Med. Chem.*, 45(12):2432–2453, 2002.

[123] E. Asante-Appiah and B. Kennedy. Protein tyrosine phosphatases: the quest for negative regulators of insulin action. *Am. J. Physiol. Endocrinol. Metab.*, 284(4):E663–E670, 2003.

[124] K. A. Kenner, E. Anyanwu, J. M. Olefsky, and J. Kusari. Protein-tyrosine phosphatase 1B is a negative regulator of insulin- and insulin-like growth factor-i-stimulated signaling. *J. Biol. Chem.*, 271(33):19810–19816, 1996.

[125] L. Klaman et al. Increased Energy Expenditure, Decreased Adiposity, and Tissue-Specific Insulin Sensitivity in Protein-Tyrosine Phosphatase 1B-Deficient Mice. *Molec. Cell. Biol.*, 20(15):5479–5489, 2000.

[126] M. Elchebly et al. Increased insulin sensitivity and obesity resistance in mice lacking the protein tyrosine phosphatase-1B gene. *Science*, 283(5407):1544–

1548, 1999.

[127] J. Liang et al. Lead identification of novel and selective TYK2 inhibitors. *Eur. J. Med. Chem.*, 67:175–187, 2013.

[128] J. Liang et al. Lead optimization of a 4-aminopyridine benzamide scaffold to identify potent, selective, and orally bioavailable TYK2 inhibitors. *J.Med. Chem.*, 56(11):4521–4536, 2013.

[129] A. Friberg et al. Discovery of potent myeloid cell leukemia 1 (MCL-1) inhibitors using fragment-based methods and structure-based design. *J. Med. Chem.*, 56(1):15–30, 2013.

[130] D. P. Wilson et al. Structure-based optimization of protein tyrosine phosphatase 1B inhibitors: From the active site to the second phosphotyrosine binding site. *J. Med. Chem.*, 50(19):4681–4698, 2007.

[131] B. Baum, M. Mohamed, M. Zayed, C. Gerlach, A. Heine, D. Hangauer, and G. Klebe. More than a simple lipophilic contact: A detailed thermodynamic analysis of nonbasic residues in the S1 pocket of thrombin. *J. Mol. Biol.*, 390(1):56–69, 2009.

[132] H. Sun, Y. Li, S. Tian, L. Xu, and T. Hou. Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set. *Phys. Chem. Chem. Phys.*, 16:16719–16729, 2014.

[133] P. A. Greenidge, C. Kramer, J. C. Mozziconacci, and R. M. Wolf. MM/GBSA binding energy prediction on the PDBbind data set: Successes, failures, and directions for further improvement. *J. Chem. Inf. Model.*, 53(1):201–209, 2013.

[134] A. Checa, A. R. Ortiz, B. De Pascual-Teresa, and F. Gago. Assessment of

solvation effects on calculated binding affinity differences: Trypsin inhibition by flavonoids as a model system for congeneric series. *J. Med. Chem.*, 40(25):4136–4145, 1997.

[135] H. Wallnoefer, K. Liedl, and T. Fox. A challenging system: Free energy prediction for factor Xa. *J. Comp. Chem.*, 32:1743–1752, 2011.

[136] T. Hou, J. Wang, Y. Li, and W. Wang. Assessing the performance of the MM/PBSA and MM/GBSA methods. I. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model*, 51(1):69–82, 2010.

[137] T. Hou, J. Wang, Y. Li, and W. Wang. Assessing the performance of the molecular mechanics/ Poisson Boltzmann surface area and molecular mechanics/ generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *J. Comp. Chem.*, 32:866–877, 2011.

[138] S. Genheden and U. Ryde. Comparison of end-point continuum-solvation methods for the calculation of protein-ligand binding free energies. *Proteins*, 80(5):1326–42, 2012.

[139] S. Genheden and U. Ryde. How to obtain statistically converged MM/GBSA results. *Journal of Computational Chemistry*, 31(4):837–846, 2010.

[140] C. M. Baker. Polarizable force fields for molecular dynamics simulations of biomolecules. *Comput. Mol. Sci.*, 5:241–254, 2015.

[141] L. Skjaerven, S. M. Hollup, and N. Reuter. Normal mode analysis for proteins. *Comput. Theor. Chem.*, 898(1-3):42–48, 2009.

[142] B. R. Miller, III, T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg. MMPBSA.py: An efficient program for end-state free energy calculations. *J. Chem. Theor. Comput.*, 8(9):3314–3321, 2012.

[143] J. Eswaran, M. Soundararajan, R. Kumar, and S. Knapp. UnPAKing the class differences among p21-activated kinases. *Trends in Biochemical Sciences*, 33:394–403, 2008.

[144] J. Qu, X. Li, B. G. Novitch, Y. Zheng, M. Kohn, J. M. Xie, S. Kozinn, R. Bronson, A. A. Beg, and A. Minden. PAK4 kinase is essential for embryonic viability and for proper neuronal development. *Molec. and Cell. Biol.*, 23:7122–7133, 2003.

[145] A. Abo, J. Qu, M. S. Cammarano, C. Dan, A. Fritsch, V. Baud, B. Belisle, and A. Minden. PAK4, a novel effector for Cdc42Hs, is implicated in the reorganization of the actin cytoskeleton and in the formation of filopodia. *The EMBO J.*, 17:6527–40, 1998.

[146] C. Dan, N. Nath, M. Liberto, and A. Minden. PAK5, a new brain-specific kinase, promotes neurite outgrowth in N1E-115 cells. *Molec. and Cell. Biol.*, 22:567–77, 2002.

[147] S. R. Lee, S. M. Ramos, A. Ko, D. Masiello, K. D. Swanson, M. L. Lu, and S. P. Balk. AR and ER interaction with a p21-activated kinase (PAK6). *Molec. Endocr.*, 16:85–99, 2002.

[148] F. Yang, X. Li, M. Sharma, M. Zarnegar, B. Lim, and Z. Sun. Androgen receptor specifically interacts with a novel p21-activated kinase, PAK6. *J. Biol. Chem.*, 276:15345–15353, 2001.

[149] Y. Baskaran, Y. W. Ng, W. Selamat, F. T. P. Ling, and E. Manser. Group I and II mammalian PAKs have different modes of activation by Cdc42. *EMBO Reports*, 13(7):653–659, 2012.

[150] B. H. Ha et al. Type II p21-activated kinases (PAKs) are regulated by an autoinhibitory pseudosubstrate. *Proc. Natl. Acad. Sci. U. S. A.*, 109(40):16107–16112, 2012.

[151] M. G. Callow, F. Clairvoyant, S. Zhu, B. Schryver, D. B. Whyte, J. R. Bischoff, B. Jallal, and T. Smeal. Requirement for PAK4 in the anchorage-independent growth of human cancer cell lines. *J. of Biol. Chem.*, 277:550, 2002.

[152] L. E. Wong, N. Chen, V. Karantza, and A. Minden. The PAK4 protein kinase is required for oncogenic transformation of MDA-MB-231 breast cancer cells. *Oncogenesis*, 2(6):e50, 2013.

[153] R. Vassar, B. D. Bennett, and others. β-secretase cleavage of alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE. *Science*, 286(5440):735–741, 1999.

[154] R. Yan, M. J. Bienkowski, et al. Membrane-anchored aspartyl protease with Alzheimer's disease β-secretase activity. *Nature*, 402(6761):533–537, 1999.

[155] S. Sinha, J. P. Anderson, et al. Purification and cloning of amyloid precursor protein β-secretase from human brain. *Nature*, 402(6761):537–540, 1999.

[156] I. Hussain, D. Powel, et al. Identification of a novel aspartic protease (asp2) as β-secretase. *Molec. and Cell. Neurosci.*, 14(6):419–427, 1999.

[157] X. Lin, G. Koelsch, et al. Human aspartic protease memapsin 2 cleaves the β-secretase site of β-amyloid precursor protein. *Proc. Nat. Acad. Sci. USA*, 97(4):1456–1460, 2000.

[158] R. E. Tanzi. The genetics of alzheimer disease. *Cold Spring Harbor Persp. Med.*, 2(10):a006296, 2012.

[159] M. Mullan, F. Crawford, et al. A pathogenic mutation for probable alzheimer's disease in the app gene at the n-terminus of β-amyloid. *Nat. Genet.*, 1(5):345–347, 1992.

[160] G. Di Fede, M. Catania, et al. A recessive mutation in the APP gene with dominant-negative effect on amyloidogenesis. *Science*, 323(5920):1473–1477, 2009.

[161] R. Vassar, D. M. Kovacs, R. Yan, and P. C. Wong. The β-secretase enzyme BACE in health and Alzheimer's disease: Regulation, cell biology, function, and therapeutic potential. *J. Neurosci.*, 29(41):12787–12794, 2009.

[162] K. D. Davies and R. C. Doebele. Molecular Pathways – ROS1 Fusion Proteins in Cancer. *Clin. Cancer Res.*, 19(15):4040–4045, 2014.

[163] A. T. Shaw et al. Crizotinib in ROS1 – Rearranged Non-Small-Cell Lung Cancer. *New Engl. J. Med.*, 371(21):1963–1971, 2014.

[164] J. J. Crawford et al. Structure-guided design of group I selective p21-activated kinase inhibitors. *J. Med. Chem.*, 58(12):5121–5136, 2015.

[165] B. W. Murray, C. Guo, J. Piraino, et al. Small-molecule p21-activated kinase inhibitor PF-3758309 is a potent inhibitor of oncogenic signaling and tumor growth. *Proc. Nat. Acad. Sci.*, 107:9446–9451, 2010.

[166] D. W. Banner et al. Mapping the conformational space accessible to BACE2 using surface mutants and cocrystals with Fab fragments, Fynomers and Xaperones. *Acta Crystallogr. D Biol. Crystallogr.*, 69(6):1124–1137, 2013.

[167] G. Tresadern et al. Rational design and synthesis of aminopiperazinones as β–secretase (bace) inhibitors. *Bioorg. Med. Chem. Lett.*, 21(24):7255–7260, 2011.

[168] I. Maffucci, A. Contini, and A. Marchesini. Explicit ligand hydration shells improve the correlation between MM-PB/GBSA binding energies and experimental activities. *J. Chem. Theor. Comput.*, 9:2706–2717, 2013.

[169] S. Genheden, T. Luchko, S. Gusarov, A. Kovalenko, and U. Ryde. An MM/3D-RISM approach for ligand binding affinities. *J. Phys. Chem.*, 114:8505–8516, 2010.

[170] A. Kohlmann, X. Zhu, and D. Dalgarno. Application of MM-GB/SA and WaterMap to SRC kinase inhibitor potency prediction. *ACS Med. Chem. Lett.*, 3(2):94–99, 2012.

[171] J. M. Heather and B. Chain. The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1–8, 2016.

[172] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, 74(12):5463–5467, 1977.

[173] T. Hunkapiller, R. J. Kaiser, B. F. Koop, and L. Hood. Large-scale and automated DNA sequence determination. *Science*, 254(5028):59–67, 1991.

[174] E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[175] J. C. Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[176] J. M. Rothberg et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475:348–352, 2011.

[177] M. Margulies and otehrs. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, 2005.

[178] K. J. McKernan et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.*, 19(9):1527–1541, 2009.

[179] D. R. Bentley et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59, 2008.

[180] Illumina Promises To Sequence Human Genome For $100 - But Not Quite Yet. https://www.forbes.com/sites/matthewherper/2017/01/09/illumina-promises-to-sequence-human-genome-for-100-but-not-quite-yet. Accessed: 30 January 2018.

[181] B. Alberts, A. Johnson, J. Lewis, M Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 4th edition, 2002.

[182] J. Quick, N. J. Loman, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530:228, 2016.

[183] M. R. Stratton, P. J. Campbell, and A. P. Futreal. The cancer genome. *Nature*, 458:719, 2009.

[184] S. Nik-Zainal et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534:47–54, 2016.

[185] E. Lumley. US Department of Energy INCITE Supercomputing Award. http://www.compbiomed.eu/department-of-energy-incite-supercomputing-award/, 2017. Accessed: 8 January 2017.

[186] Qatar Cancer Society. Breast Cancer. http://www.qcs.qa/. Accessed: 30 January 2018.

[187] S. W. Fanning et al. Estrogen receptor alpha somatic mutations Y537S and D538G confer breast cancer endocrine resistance by stabilizing the activating function-2 binding conformation. *eLife*, pages 1–25, 2016.

[188] A. M. Brzozowski et al. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature*, 389(6652):753–758, 1997.

[189] J. A. Katzenellenbogen et al. Tripartite steroid hormone receptor pharmacology: Interaction with multiple effector sites as a basis for the cell- and promoter- specific action of these hormones. *Mol. Endocrinol.*, 10(2):119–131, 1996.

[190] M. Beato et al. Interaction of steroid hormone receptors with the transcription initiation complex. *Endocrinol. Rev.*, 7(6):587–609, 1996.

[191] M. J. Tsai and O'Malley B. W. Molecular mechanisms of action of steroid/thyroid receptor superfamily members. *Annu. Rev. Biochem.*, 63:451–486, 1994.

[192] M. Beato et al. Steroid hormone receptors: Many actors in search of a plot. *Cell Rev.*, 83:851–857, 1995.

[193] J. A. Veltman and H. G. Brunner. De novo mutations in human genetic disease. *Nat. Rev. Genet.*, 13:565, 2012.

[194] A. E. Fliss, S. Benzeno, J. Rao, and A. J. Caplan. Control of estrogen receptor ligand binding by Hsp90. *J. Steroid Biochem. Mol. Biol.*, 72(5):223–230, 2000.

[195] J. S. Lewis and V. C. Jordan. Selective estrogen receptor modulators (SERMs): Mechanisms of anticarcinogenesis and drug resistance. *Mut. Res.*, 591(1-2):247–263, 2005.

[196] P. M. Henttu, E. Kalkhoven, and M. G. Parker. AF-2 activity and recruitment of steroid receptor coactivator 1 to the estrogen receptor depend on a lysine residue conserved in nuclear receptors. *Mol. Cell. Biol.*, 17(4):1832–9, 1997.

[197] L. Liao, S. Q. Kuang, Y. Yuan, S. M. Gonzalez, B. W. O'Malley, and J. Xu. Molecular structure and biological function of the cancer-amplified nuclear receptor coactivator SRC-3/AIB1. *J. Steroid Biochem. Mol. Biol.*, 83(1):3–14, 2002.

[198] K. E. Carlson, I. Choi, A. Gee, B. S. Katzenellenbogen, and J. A. Katzenellenbogen. Altered Ligand Binding Properties and Enhanced Stability of a Constitutively Active Estrogen Receptor: Evidence That an Open Pocket Conformation Is Required for Ligand Interaction. *Biochem.*, 36(48):14897–14905, 1997.

[199] A. K. Shiau et al. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, 95(7):927–937, 1998.

[200] M. Gangloff, M. Ruff, S. Eiler, S. Duclaud, J. M. Wurtz, and D. Moras. Crystal structure of a mutant herα ligand-binding domain reveals key structural features for the mechanism of partial agonism. *J. Biol. Chem.*, 276(18):15059–15065, 2001.

[201] S. Wan, A. P. Bhati, S. J. Zasada, I. Wall, D. Green, and P. V. Coveney. Rapid and reliable binding affinity prediction for analysis of bromodomain inhibitors : a computational study. *J. Chem. Theor. Comp.*, pages 1–30, 2017.

[202] A. Jakalian, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comp. Chem.*, 23(16):1623–1641, 2002.

[203] S. Wan, A. P. Bhati, S. Skerratt, K. Omoto, V. Shanmugasundaram, S. K. Bagal, and P. V. Coveney. Evaluation and Characterization of Trk Kinase Inhibitors for the Treatment of Pain: Reliable Binding Affinity Predictions from Theory and Computation. *J. Chem. Inf. Model.*, 57(4):897–909, 2017.